Chapter 2
Defining the Model and Parameter

Sherri Rose, Mark J. van der Laan

Targeted statistical learning from data is often concerned with the estimation of causal effects and an assessment of uncertainty for the estimator. In Chap. 1, we identified the road map we will follow to solve this estimation problem. Now, we formalize the concepts of the model and target parameter. We will introduce additional topics that may seem abstract. While we attempt to elucidate these abstractions with tangible examples, depending on your background, the material may be quite dense compared to other textbooks you have read. Do not get discouraged. Sometimes a second reading and careful notes are helpful and sufficient to illuminate these concepts. Researchers and students at UC Berkeley have also had great success discussing these topics in groups. If this is your assigned text for a course or workshop, meet outside of class with your fellow classmates. We guarantee you that the effort is worth it so you can move on to the next step in the targeted learning road map. Once you have a firm understanding of the core material in Chap. 2, you can begin the estimation steps.

This chapter is based on methods pioneered by Judea Pearl, and we consider his text *Causality*, recently published in a second edition (Pearl 2009), a companion book to our book. Causal inference requires both a causal model to define the causal effect as a target parameter of the distribution of the data *and* robust semiparametric efficient estimation, with his book covering the former and ours the latter. We start by succinctly summarizing the open problem:

> The statistical estimation problem begins by defining a statistical model $\mathcal{M}$ for P_0. The statistical model $\mathcal{M}$ is a collection of possible probability distributions P of O. P_0 is the true distribution of O. The estimation problem requires the description of a target parameter of P_0 one wishes to learn from the data. This definition of a target parameter requires specification of a mapping Ψ one can then apply to P_0. Clearly, this mapping Ψ needs to be defined on any possible probability distribution in the statistical model $\mathcal{M}$. Thus Ψ maps any $P \in \mathcal{M}$ into a vector of numbers $\Psi(P)$. We write the mapping as $\Psi : \mathcal{M} \to \mathbb{R}^d$ for a

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_2,

d-dimensional parameter. We introduce ψ_0 as the evaluation of $\Psi(P_0)$, i.e., the true value of our parameter. The statistical estimation problem is now to map the observed data $O_1, \ldots, O_n$ into an estimator of $\Psi(P_0)$ that incorporates the knowledge that $P_0 \in \mathcal{M}$, accompanied by an assessment of the uncertainty in the estimator.

In the following sections, we will define a model that goes beyond a statistical model by incorporating nontestable assumptions, define a parameter of interest in that model that can be interpreted as a causal effect, determine the assumptions to establish the identifiability of the causal parameter from the distribution of the observed data, and, finally, based on this modeling and identifiability exercise, commit to a statistical model (i.e., $\mathcal{M}$) and target parameter (i.e., Ψ).

Recall that the data $O_1, \ldots, O_n$ consist of n i.i.d. copies of a random variable O with probability distribution P_0. For a data structure, such as $O = (W, A, Y)$ with covariates W, exposure A, and outcome Y discrete, which we use as a simple example in this chapter, uppercase letters represent random variables and lowercase letters are a specific value for that variable. For example, if all variables are discrete, $P_0(W = w, A = a, Y = y)$ assigns a probability to any possible outcome (w, a, y) for $O = (W, A, Y)$.

2.1 Defining the Structural Causal Model

We first specify a set of endogenous variables $X = (X_j : j)$. Endogenous variables are those variables for which the structural causal model (SCM) will state that it is a (typically unknown) deterministic function of some of the other endogenous variables and an exogenous error. Typically, the endogenous variables X include the observables O, but might also include some nonobservables that are meaningful and important to the scientific question of interest. Perhaps there was a variable you did not measure, but would have liked to, and it plays a crucial role in defining the scientific question of interest. This variable would then be an unobserved endogenous variable. For example, if you are studying the effect of hepatitis B on liver cancer, you might also want to measure hepatitis C and aflatoxin exposure. However, suppose you know the role aflatoxin plays in the relationships between hepatitis B and liver cancer, but you were unable to measure it. Aflatoxin exposure is, therefore, an unobserved endogenous variable. Liver cancer, hepatitis B, and hepatitis C are observed endogenous variables.

In a very simple example, we might have $j = 1, \ldots, J$, where $J = 3$. Thus, $X = (X_1, X_2, X_3)$. We can rewrite X as $X = (W, A, Y)$ if we say $X_1 = W$, $X_2 = A$, and $X_3 = Y$. Let W represent the set of baseline covariates for a subject, A the treatment or exposure, and Y the outcome. All the variables in X are observed. Suppose we are interested in estimating the effect of leisure-time physical activity (LTPA) on mortality in an elderly population. A study is conducted to estimate this effect where

we sample individuals from the population of interest. The hypothesis is that LTPA at or above current recommended levels decreases mortality risk. Let us say that LTPA is a binary variable $A \in \{0, 1\}$ defined by the recommended level of energy expenditure. For all subjects meeting this level, $A = 1$ and all those below have $A = 0$. The mortality outcome is also binary $Y \in \{0, 1\}$ and defined as death within 5 years of the beginning of the study, with $Y = 1$ indicating death. W includes variables such as age, sex, and health history.

For each endogenous variable X_j one specifies the parents of X_j among X, denoted $Pa(X_j)$. In our mortality study example above, the parent of A is the set of baseline covariates W. Thus, $Pa(A) = W$. The specification of the parents might be known by the time ordering in which the X_j were collected over time: the parents of a variable collected at time t could be defined as the observed past at time t. This is true for our study of LTPA; $W = \{\text{age, sex, health history}\}$ all occur before the single measurement of LTPA. Likewise, LTPA was generated after the baseline covariates and before death but depends on the baseline covariates. Death was generated last and depends on both LTPA and the baseline covariates. We can see the time ordering involved in this process: the baseline covariates occurred before the exposure LTPA, which occurred before the outcome of death: $W \rightarrow A \rightarrow Y$.

We denote a collection of exogenous variables by $U = (U_{X_j} : j)$. These variables in U are never observed and are not affected by the endogenous variables in the model, but instead they affect the endogenous variables. They may also be referred to as background or error variables. One assumes that X_j is some function of $Pa(X_j)$ and an exogenous U_{X_j}:

$$X_j = f_{X_j}(Pa(X_j), U_{X_j}), \ j = 1 \ldots, J.$$

The collection of functions f_{X_j} indexed by all the endogenous variables is represented by $f = (f_{X_j} : j)$. Together with the joint distribution of U, these functions f_{X_j}, specify the data-generating distribution of (U, X) as they describe a deterministic system of structural equations (one for each endogenous variable X_j) that deterministically maps a realization of U into a realization of X. In an SCM one also refers to some of the endogenous variables as intervention variables. The SCM assumes that intervening on one of the intervention variables by setting their value, thereby making the function for that variable obsolete, does not change the form of the other functions. The functions f_{X_j} are often unspecified, but in some cases it might be reasonable to assume that these functions have to fall in a certain more restrictive class of functions. Similarly, there might be some knowledge about the joint distribution of U. The set of possible data-generating distributions of (U, X) can be obtained by varying the structural equations f over all allowed forms, and the distribution of the errors U over all possible error distributions defines the SCM for the full-data (U, X), i.e., the SCM is a statistical model for the random variable (U, X). An example of a fully parametric SCM would be obtained by assuming that all the functions f_{X_j} are known up to a finite number of parameters and that the error distribution is a multivariate normal distribution with mean zero and unknown covariance matrix. Such

parametric structural equation models are not recommended, for the same reasons as outlined in Chap. 1.

The corresponding SCM for the observed data O also includes specifying the relation between the random variable (U, X) and the observed data O, so that the SCM for the full data implies a parameterization of the probability distribution of O in terms of f and the distribution P_U of U. This SCM for the observed data also implies a statistical model for the probability distribution of O.

Let's translate these concepts into our mortality study example. We have the functions $f = (f_W, f_A, f_Y)$ and the exogenous variables $U = (U_W, U_A, U_Y)$. The values of W, A, and Y are deterministically assigned by U corresponding to the functions f. We specify our structural equation models, based on investigator knowledge, as

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(W, U_A), \\ Y &= f_Y(W, A, U_Y), \end{aligned} \tag{2.1}$$

where no assumptions are made about the true shape of f_W, f_A, and f_Y. These functions f are nonparametric as we have not put a priori restrictions on their functional form. We may assume that U_A is independent of U_Y, given W, which corresponds with believing that there are no unmeasured factors that predict both A and the outcome Y: this is often called the no unmeasured confounders assumption. This SCM represents a semiparametric statistical model for the probability distribution of the errors U and endogenous variables $X = (W, A, Y)$. We assume that the observed data structure $O = (W, A, Y)$ is actually a realization of the endogenous variables (W, A, Y) generated by this system of structural equations. This now defines the SCM for the observed data O. It is easily seen that any probability distribution of O can be obtained by selecting a particular data-generating distribution of (U, X) in this SCM. Thus, the statistical model for P_0 implied by this SCM is a nonparametric model. As a consequence, one cannot determine from observing O if the assumptions in the SCM contradict the data. One states that the SCM represents a set of nontestable causal assumptions we have made about how the data were generated in nature.

Specifically, with the SCM represented in (2.1), we have assumed that the underlying data were generated by the following actions:

1. Drawing unobservable U from some probability distribution P_U ensuring that U_A is independent of U_Y, given W,
2. Generating W as a deterministic function of U_W,
3. Generating A as a deterministic function of W and U_A,
4. Generating Y as a deterministic function of W, A, and U_Y.

What if, instead, our SCM had been specified as follows:

$$\begin{aligned} W &= f_W(U_W), \\ A &= f_A(U_A), \\ Y &= f_Y(W, A, U_Y). \end{aligned} \tag{2.2}$$

What different assumption are we making here? If you compare (2.1) and (2.2), you see that the only difference between the two is the structural equation for f_A. In (2.2), A is evaluated as a deterministic function of U_A only. The baseline variables W play no role in the generation of variable A. We say that (2.2) is a more restrictive SCM than (2.1) because of this additional assumption about data generation. When might a researcher make such an assumption? In Chap. 1, we discussed RCTs. RCTs are studies where the subjects are randomized to treatment in the study. If our study of LTPA had been an RCT, it would make sense to assume the SCM specified in (2.2) given our knowledge of the study design. However, since it would be unethical to randomize subjects to levels of exercise, given the known health benefits, our study of LTPA on mortality is observational and we assume the less restrictive (2.1).

Causal assumptions made by the SCM for the full data:

- For each endogenous X_j, $X_j = f_j(Pa(X_j), U_{X_j})$ only depends on the other endogenous variables through its parents $Pa(X_j)$.
- The exogenous variables have a particular joint distribution P_U.

The SCM for the observed data includes the following additional assumption:

- The probability distribution of observed data structure O is implied by the probability distribution of (U, X).

After having specified the parent sets $Pa(X_j)$ for each endogenous variable X_j, one might make an assumption about the joint distribution of U, denoted P_U, representing knowledge about the underlying random variable (U, X) as accurately as possible. This kind of assumption would typically not put any restrictions on the probability distribution of O. The underlying data (U, X) are comprised of the exogenous variables U and the endogenous variables X, which is why we use the notation (U, X). In a typical SCM, the endogenous variables are the variables for which we have some understanding, mostly or fully observed, often collected according to a time ordering, and are very meaningful to the investigator. On the other hand, typically much of the distribution of U is poorly understood. In particular, one would often define U_{X_j} as some surrogate of potential unmeasured confounders, collapsing different poorly understood phenomena in the real world in one variable. The latter is reflected by the fact that we do not even measure these confounders, or know how to measure them. However, in some applications something about the joint distribution of U might be understood, and some components of U might be measured. For example, it might be known that treatment was randomized as in an RCT, implying that the error U_A for that treatment variable is independent of all other errors. On the other hand, in an observational study, one might feel uncomfortable making the assumption that U_A is independent of U_Y, given W, since one might know that some of the true confounders were not measured and are thereby captured by U_A.

Relationship of X and O. Our observed random variable O is related to X, and has a probability distribution that is implied by the distribution of (U, X). Specification of this relation is an important assumption of the SCM for the observed data O. A typical example is that $O = \Phi(X)$ for some Φ, i.e., O is a function of X. This includes the special case that $O \subset X$, i.e., with O being a simple subset of X. Because of this relationship $O = \Phi(X)$, the marginal probability distribution of X,

$$P_X(x) = \sum_u P_f(X = x \mid U = u) P_U(U = u),$$

also identifies the probability distribution of O through the functions $f = (f_{X_j} : j)$ and the distribution of the exogenous errors U. [Note that the conditional probability distribution $P_f(X = x \mid U = u)$ of X, given a realization $U = u$, is indeed completely determined by the functions f, which explains our notation P_f.] For example, if $X = O$, then:

$$P(o) = \sum_u P_f(X = o \mid U = u) P_U(U = u).$$

In order to make explicit that the probability distribution P of O is implied by the probability distribution of (U, X), we use the notation $P = P(P_{U,X})$. The true probability distribution $P_{U,X,0}$ of (U, X) implies the true probability distribution P_0 of O through this relation: $P_0 = P(P_{U,X,0})$. Since the assumed SCM often does not put any restrictions on the functions f_{X_j}, and the selection of the parent sets $Pa(X_j)$ might be purely based on time ordering (thereby not implying conditional independencies among the X_js), for many types of restrictions one would put on P_U, the resulting SCM for (U, X) would still not provide any restriction on the distribution of O. In that case, these causal assumptions provide no restriction on the distribution of O itself and thus imply a nonparametric *statistical* model $\mathcal{M}$ for the distribution P_0 of O. This statistical model $\mathcal{M}$ implied by the SCM for the observed data is given by $\mathcal{M} = \{P(P_{U,X}) : P_{U,X}\}$, where $P_{U,X}$ varies over all possible probability distributions of (U, X) in the SCM.

Each possible probability distribution $P_{U,X}$ of (U, X) in the SCM for the full data, indexed by a choice of error distribution P_U and a set of deterministic functions $(f_{X_j} : j)$, implies a probability distribution $P(P_{U,X})$ of O. In this manner the SCM for the full data implies a parameterization of the true probability distribution of O in terms of a true probability distribution of (U, X), so that the statistical model $\mathcal{M}$ for the probability distribution P_0 of O can be represented as $\mathcal{M} = \{P(P_{U,X}) : P_{U,X}\}$, where $P_{U,X}$ varies over all allowed probability distributions of (U, X) in the SCM. If this statistical model $\mathcal{M}$ implied by the SCM is nonparametric, then it follows that none of the causal assumptions encoded by the SCM are testable from the observed data.

2.2 Causal Graphs

SCMs provide a system for assigning values to a set of variables from random input. They are also an effective and straightforward means for explicitly specifying causal assumptions and the identifiability of the causal parameter of interest based on the observed data. We can draw a causal graph from our SCM, which is a visual way to describe some of the assumptions made by the model and the restrictions placed on the joint distribution of the data (U, X). However, in this text we do not place heavy emphasis on causal graphs as their utility is limited in many situations (e.g., complicated longitudinal data structures), and simpler visual displays of time ordering may provide more insight. Causal graphs also cannot encode every assumption we make in our SCM, and, in particular, the identifiability assumptions derived from causal graphs alone are not specific for the causal parameter of interest. Identifiability assumptions derived from a causal graph will thus typically be stronger than required. In addition, the link between the observed data and the full-data model represented by the causal graph is often different than simply stating that O corresponds with observing a subset of all the nodes in the causal graph. In this case, the causal graph itself cannot be used to assess the identifiability of a desired causal parameter from the observed data distribution.

2.2.1 Terminology

Figure 2.1 displays a possible causal graph for (2.1). The graph is drawn based on the relationships defined in f. The parents $Pa(X_j)$ of each X_j are connected to each X_j with an arrow directed toward X_j. Each X_j also has a directed arrow connecting its U_{X_j}. For example, the parents of Y, those variables in X on the right-hand side of the equation f_Y, are A and W. In Fig. 2.1, A and W are connected to Y, the child, with directed arrows, as is the exogenous U_Y. The baseline covariates W are represented with one variable. All the variables X and U in the graph are called nodes, and the lines that connect nodes are edges. All ancestors of a node occur before that node and all descendants occur after that node. This is a directed graph, meaning that each edge has only one arrow.

A path is any sequence of edges in a graph connecting two nodes. An example of a directed path in Fig. 2.1 is $W \rightarrow A \rightarrow Y$. This path connects each node with arrows that point in the direction of the path. In this figure there are several backdoor paths, which are paths that start with a node that has a directed arrow pointing into that node. The path can then be followed without respect to the direction of the arrows. For example, the path from Y to A through W is a backdoor path. Likewise, the path from Y to W through A is a backdoor path. These graphs are also acyclic; you cannot start at a node in a directed path and then return back to the same node through a closed loop. A collider is a node in a path where both arrows are directed toward the node. There are no colliders in Fig. 2.1. A blocked path is any path with at least one collider. A direct effect is illustrated by a directed arrow between two nodes, with

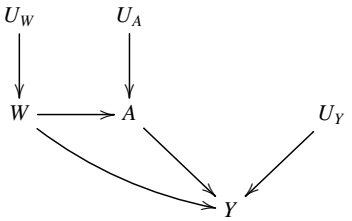

Fig. 2.1 A possible causal graph for (2.1).

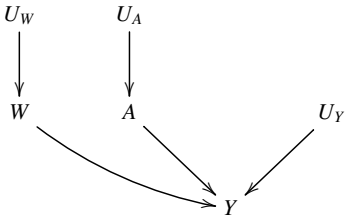

Fig. 2.2 A possible causal graph for (2.2)

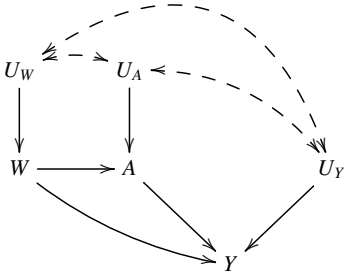

Fig. 2.3 A causal graph for (2.1) with no assumptions on the distribution of P_U

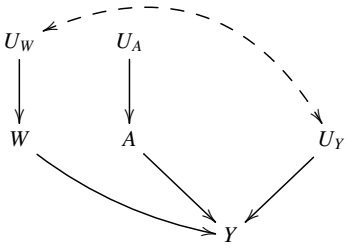

Fig. 2.4 A causal graph for (2.2) with no assumptions on the relationship between U_W and U_Y

no nodes mediating the path. Any unblocked path from A to Y other than the direct effect connecting A and Y represents an indirect effect of A on Y. One must block all unblocked backdoor paths from A to Y in order to isolate the causal effect of A on Y.

2.2.2 Assumptions

In Sect. 2.1, we discussed the typically nontestable causal assumptions made by an SCM. We make the first assumption by defining the parents $Pa(X_j)$ for each endogenous X_j. The second is any set of assumptions about the joint distribution P_U of the exogenous variables.

The assumptions made based on actual knowledge concerning the relationships between variables [i.e., defining the parents $Pa(X_j)$ for each endogenous X_j] are displayed in our causal graph through the presence and absence of directed arrows. The explicit absence of an arrow indicates a known lack of a direct effect. In many cases all arrows are included as it is not possible to exclude a direct effect based on a priori knowledge. In Fig. 2.1, the direction of the arrows is defined by the assignment of the parents to each node, including the time ordering assumed during the specification of (2.1). There is no explicit absence of any arrows; no direct effects are excluded. However, if we were to draw a graph for (2.2), it could look like Fig. 2.2. The direct effect between W and A is excluded because A is evaluated as a deterministic function of U_A only.

The assumptions on the distribution P_U are reflected in causal graphs through dashed double-headed arrows between the variables U. In Figs. 2.1 and 2.2, there are no arrows between the $U = (U_W, U_A, U_Y)$. Therefore, (2.1) and (2.2) included the assumption of joint independence of the endogenous variables U, which is graphically displayed by the lack of arrows. This is not an assumption one is usually able to make based on actual knowledge. More likely, we are able to make few or no assumptions about the distribution of P_U.

For (2.1), with no assumptions about the distribution of P_U, our causal graph would appear as in Fig. 2.3. For (2.2), our causal graph based on actual knowledge may look like Fig. 2.4. Since A is randomized, this implies that U_A is independent of U_Y and U_W, and we remove the arrows connecting U_A to U_Y and U_A to U_W. However, we have no knowledge to indicate the independence of U_Y and U_W, thus we cannot remove the arrows between these two variables.

The causal graph encodes some of the information and assumptions described by the SCM. It is an additional tool to visually describe assumptions encoded by the SCM. In more complex longitudinal data structures, it may be simpler to work with the SCM over the causal graph, as the intricacies of the causal relationships and abundance of arrows can limit the utility of the graphic.

2.3 Defining the Causal Target Parameter

Now that we have a way of modeling the data-generating mechanism with an SCM, we can focus on what we are trying to learn from the observed data. That is, we can define a causal target parameter of interest as a parameter of the distribution of the full-data (U, X) in the SCM. Formally, we denote the SCM for the full-data (U, X) by $\mathcal{M}^F$, a collection of possible $P_{U,X}$ as described by the SCM. In other words, $\mathcal{M}^F$, a model for the full data, is a collection of possible distributions for the underlying data (U, X). Ψ^F is a mapping applied to a $P_{U,X}$ giving $\Psi^F(P_{U,X})$ as the target parameter of $P_{U,X}$. This mapping needs to be defined for each $P_{U,X}$ that is a possible distribution of (U, X), given our assumptions coded by the posed SCM. In this way, we state $\Psi^F : \mathcal{M}^F \rightarrow \mathbb{R}^d$, where $\mathbb{R}^d$ indicates that our parameter is a vector of d real numbers. The SCM $\mathcal{M}^F$ consists of the distributions indexed by the deterministic function $f = (f_{X_j} : j)$ and distribution P_U of U, where f and this joint distribution P_U are identifiable from the distribution of the full-data (U, X). Thus the target parameter can also be represented as a function of f and the joint distribution of U.

Recall our mortality example with data structure $O = (W, A, Y)$ and SCM given in (2.1) with no assumptions about the distribution P_U. We can define $Y_a = f_Y(W, a, U_Y)$ as a random variable corresponding with intervention $A = a$ in the SCM. The marginal probability distribution of Y_a is thus given by

$$P_{U,X}(Y_a = y) = P_{U,X}(f_Y(W, a, U_Y) = y).$$

The causal effect of interest for a binary A (suppose it is the causal risk difference) could then be defined as a parameter of the distribution of (U, X) given by

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

In other words, $\Psi^F(P_{U,X})$ is the difference of marginal means of counterfactuals Y_1 and Y_0. We discuss this in more detail in the next subsection.

2.3.1 Interventions

We will define our causal target parameter as a parameter of the distribution of the data (U, X) under an intervention on one or more of the structural equations in f. The intervention defines a random variable that is a function of (U, X), so that the target parameter is $\Psi^F(P_{U,X})$. In Chap. 1, we discussed the "ideal experiment" which we cannot conduct in practice, where we observe each subject's outcome at all levels of A under identical conditions. Intervening on the system defined by our SCM describes the data that would be generated from the system at the different levels of our intervention variable (or variables). For example, in our study of LTPA on mortality, we can intervene on the exposure LTPA in order to observe the results

of this intervention on the system. By assumption, intervening and changing the functions f_{X_j} of the intervention variables does not change the other functions in f. With the SCM given in (2.1) we can intervene on f_A and set $a = 1$:

$$\begin{aligned} W &= f_W(U_W), \\ a &= 1, \\ Y_1 &= f_Y(W, 1, U_Y). \end{aligned}$$

We can also intervene and set $a = 0$:

$$\begin{aligned} W &= f_W(U_W), \\ a &= 0, \\ Y_0 &= f_Y(W, 0, U_Y). \end{aligned}$$

The intervention defines a random variable that is a function of (U, X), namely, $Y_a = Y_a(U)$ for $a = 1$ and $a = 0$. The notation $Y_a(U)$ makes explicit that Y_a is random only through U. The probability distribution of the (X, U) under an intervention is called the postintervention distribution. Our target parameter is a parameter of the postintervention distribution of Y_0 and Y_1, i.e., it is a function of these two postintervention distributions, namely, some difference. Thus, the SCM for the full data allows us to define the random variable $Y_a = f_Y(W, a, U_Y)$ for each a, where Y_a represents the outcome that would have been observed under this system for a particular subject under exposure a. Thus, with the SCM we can carry out the "ideal experiment" and define parameters of the distribution of the data generated in this perfect experiment, even though our observed data are only the random variables $O_1, \ldots, O_n$.

Formally, and more generally, the definition of the target parameter involves first specifying a subset of the endogenous nodes X_j playing the role of intervention nodes. Let A_s denote the intervention nodes, $s = 0, \ldots, S$, so that $A = (A_s : s = 1, \ldots, S)$, which, in shorthand notation, we also denote by $A = (A_s : s)$. We will denote the other endogenous nodes in X by $L = (L_r : r)$. Thus, $X = ((A_s : s), (L_r : r))$. Static interventions on the A-nodes correspond with setting A to a fixed value a, while dynamic interventions deterministically set A_s according to a fixed rule applied to the parents of A_s. Static interventions are a subset of the dynamic interventions. We will denote such a rule for assigning d to the intervention nodes, but it should be observed that d defines a rule for each A_s. Thus $d = (d_s : s = 1, \ldots, S)$ is a set of S rules. Such rules d are also called dynamic treatment regimens.

For a particular intervention d on the A nodes, and for a given realization u, the SCM generates deterministically a corresponding value for L, obtained by erasing the f_{A_s} functions, and carrying out the intervention d on A in the parent sets of the remaining equations. We denote the resulting realization by $L_d(u)$ and note that $L_d(u)$ is implied by f and u. The actual random variable $L_d(U)$ is called a postintervention random variable corresponding with the intervention that assigns the intervention nodes according to rule d. The probability distribution of $L_d(U)$ can be described as

$$P(L_d(U) = l) = \sum_u P_f(L_d(u) = l \mid U = u)P_U(u) = \sum_u I(L_d(u) = l)P_U(u).$$

In other words, it is the probability that U falls in the set of u-realizations under which the SCM system deterministically sets $L_d(u) = l$. Indicator $I(L_d(u) = l)$ is uniquely determined by the function specifications f_{X_j} for the X_j nodes that comprise L. This shows explicitly that the distribution of $L_d(U)$ is a parameter of f and the distribution of U, and thus a well-defined parameter on the full-data SCM $\mathcal{M}^F$ for the distribution of (U, X). We now define our target parameter $\Psi^F(P_{U,X})$ as some function of $(P_{L_d} : d)$ for a set of interventions d. Typically, we define our target parameter as a so-called causal contrast that involves a difference between two of such d-specific postintervention probability distributions. This target parameter is referred to as a causal parameter since it is a parameter of the postintervention distribution of L as a function of an intervention choice on $A = (A_s : s)$ across one or more interventions.

2.3.2 Counterfactuals

We would ideally like to see each individual's outcome at all possible levels of exposure A. The study is only capable of collecting Y under one exposure, the exposure the subject experiences. We discussed interventions on our SCM in Sect. 2.3.1 and we intervened on A to set $a = 1$ and $a = 0$ in order to generate the outcome for each subject under $A = a$ in our mortality study. Recall that Y_a represents the outcome that would have been observed under this system for a particular subject under exposure a. For our binary exposure LTPA , we have $(Y_a : a)$, with $a \in \mathcal{A}$, and where $\mathcal{A}$ is the set of possible values for our exposure LTPA. Here, this set is simply $\{0, 1\}$, but in other examples it could be continuous or otherwise more complex. Thus, in our example, for each realization u, which might correspond with an individual randomly drawn from some target population, by intervening on (2.1), we can generate so-called counterfactual outcomes $Y_1(u)$ and $Y_0(u)$. These counterfactual outcomes are implied by our SCM; they are consequences of it. That is, $Y_0(u) = f_Y(W, 0, u_Y)$, and $Y_1(u) = f_Y(W, 1, u_Y)$, where $W = f_W(u_W)$ is also implied by u. The random counterfactuals $Y_0 = Y_0(U)$ and $Y_1 = Y_1(U)$ are random through the probability distribution of U. Now we have the expected outcome had everyone in the target population met or exceeded recommended levels of LTPA, and the expected outcome had everyone had levels of LTPA below health recommendations. For example, the expected outcome of Y_1 is the mean of $Y_1(u)$ with respect to the probability distribution of U. Our target parameter is a function of the probability distributions of these counterfactuals: $E_0Y_1 - E_0Y_0$.

2.3.3 Establishing Identifiability

Are the assumptions we have already made enough to express the causal parameter of interest as a parameter of the probability distribution P_0 of the observed data? We want to be able to write $\Psi^F(P_{U,X,0})$ as $\Psi(P_0)$ for some parameter mapping Ψ, where we remind the reader that the SCM also specifies how the distribution P_0 of the observed data structure O is implied by the true distribution $P_{U,X,0}$ of (U, X). Since the true probability distribution of (U, X) can be any element in the SCM $\mathcal{M}^F$, and each such choice $P_{U,X}$ implies a probability distribution $P(P_{U,X})$ of O, this requires that we show that $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$ for all $P_{U,X} \in \mathcal{M}^F$.

This step involves establishing possible additional assumptions on the distribution of U, or sometimes also on the deterministic functions f, so that we can identify the target parameter from the observed data distribution. Thus, for each probability distribution of the underlying data (U, X) satisfying the SCM with these possible additional assumptions on P_U, we have $\Psi^F(P_{U,X}) = \Psi(P(P_{U,X}))$ for some Ψ. O is implied by the distribution of (U, X), such as $O = X$ or $O \subset X$, and $P = P(P_{X,U})$, where $P(P_{U,X})$ is a distribution of O implied by $P_{U,X}$.

Let us denote the resulting full-data SCM by $\mathcal{M}^{F*} \subset \mathcal{M}^F$ to make clear that possible additional assumptions were made that were driven purely by the identifiability problem, not necessarily reflecting reality. To be explicit, $\mathcal{M}^F$ is the full-data SCM under the assumptions based on real knowledge, and $\mathcal{M}^{F*}$ is the full-data SCM under possible additional causal assumptions required for the identifiability of our target parameter. We now have that for each $P_{U,X} \in \mathcal{M}^{F*}$, $\Psi^F(P_{U,X}) = \Psi(P)$, with $P = P(P_{U,X})$ the distribution of O implied by $P_{U,X}$ (whereas P_0 is the true distribution of O implied by the true distribution $P_{U,X,0}$).

Theorems exist that are helpful to establish such a desired identifiability result. For example, if $O = X$, and the distribution of U is such that, for each s, A_s is independent of L_d, given $Pa(A_s)$, then the well-known g-formula expresses the distribution of L_d in terms of the distribution of O:

$$P(L_d = l) = \prod_{r=1}^{R} P(L_r = l_r \mid Pa_d(L_r)) = Pa_d(l_r)),$$

where $Pa_d(L_r)$ are the parents of L_r with the intervention nodes among these parent nodes deterministically set by intervention d.

This so-called sequential randomization assumption can be established for a particular independence structure of U by verifying the backdoor path criterion on the corresponding causal graph implied by the SCM and this independence structure on U. The backdoor path criterion states that for each A_s, each backdoor path from A_s to an L_r node that is realized after A_s is blocked by one of the other L_r nodes.

In this manner, one might be able to generate a number of independence structures on the distribution of U that provide the desired identifiability result. That is, the resulting model for U that provides the desired identifiability might be represented as a union of models for U that assume a specific independence structure.

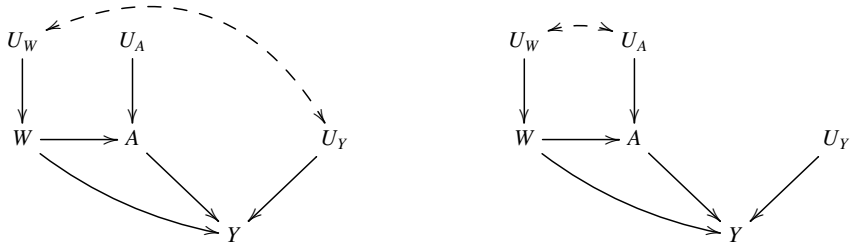

Fig. 2.5 Causal graphs for (2.1) with various assumptions about the distribution of P_U

If there is only one intervention node, i.e., $S = 1$, so that $O = (W, A, Y)$, the sequential randomization assumption reduces to the randomization assumption. The randomization assumption states that treatment node A is independent of counterfactual Y_a, conditional on W: $Y_a \perp A \mid Pa(A) = W$. You may be familiar with the (sequential) randomization assumption by another name, the no unmeasured confounders assumption. For our purposes, confounders are those variables in X one needs to observe in O in order to establish the identifiability of the target parameter of interest. We note that different such subsets of X may provide a desired identifiability result.

If we return to our mortality example and the structural equation models found in (2.1), the union of several independence structures allows for the identifiability of our causal target parameter $E_0 Y_1 - E_0 Y_0$ by meeting the backdoor path criterion. The independence structure in Fig. 2.3 does not meet the backdoor path criterion, but the two in Fig. 2.5 do. Thus in these two graphs the randomization assumption holds: A and Y_a are conditionally independent given W, which is implied by U_A being independent of U_Y, given W. It should be noted that Fig. 2.1 is a special case of the first graph in Fig. 2.5, so the union model for the distribution of U only represents two conditional independence models.

2.3.4 Commit to a Statistical Model and Target Parameter

The identifiability result provides us with a purely statistical target parameter $\Psi(P_0)$ on the distribution P_0 of O. The full-data model $\mathcal{M}^{F*}$ implies a statistical observed data model $\mathcal{M} = \{P(P_{X,U}) : P_{X,U} \in \mathcal{M}^{F*}\}$ for the distribution $P_0 = P(P_{U,X,0})$ of O. This now defines a target parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$. The statistical observed data model for the distribution of O might be the same for $\mathcal{M}^F$ and $\mathcal{M}^{F*}$. If not, then one might consider extending the Ψ to the larger statistical observed data model implied by $\mathcal{M}^F$, such as possibly a fully nonparametric model allowing for all probability distributions. In this way, if the more restricted SCM holds, our target parameter would still estimate the target parameter, but one now also allows the data to contradict the more restricted SCM based on additional doubtful assumptions.

We can return to our example of the effect of LTPA on mortality and define our parameter, the causal risk difference, in terms of the corresponding statistical parameter $\Psi(P_0)$:

$$\Psi^F(P_{U,X,0}) = E_0Y_1 - E_0Y_0 = E_0[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)] \equiv \Psi(P_0),$$

where the outer expectation in the definition of $\Psi(P_0)$ is the mean across the strata for W. This identifiability result for the additive causal effect as a parameter of the distribution P_0 of O required making the randomization assumption stating that A is independent of the counterfactuals (Y_0, Y_1) within strata of W. This assumption might have been included in the original SCM $\mathcal{M}^F$, but, if one knows there are unmeasured confounders, then the model $\mathcal{M}^{F*}$ would be more restrictive by enforcing this "known to be wrong" randomization assumption.

Another required assumption is that $P_0(A = 1, W = w) > 0$ and $P_0(A = 0, W = w) > 0$ are positive for each possible realization w of W. Without this assumption, the conditional expectations of Y in $\Psi(P_0)$ are not well defined. This positivity assumption is often called the experimental treatment assignment (ETA) assumption. Here we are assuming that the conditional treatment assignment probabilities are positive for each possible w: $P_0(A = 1 \mid W = w) > 0$ and $P_0(A = 0 \mid W = w) > 0$ for each possible w. However, the positivity assumption is a more general name for the condition that is necessary for the target parameter $\Psi(P_0)$ to be well defined, and it often requires the censoring or treatment mechanism to have certain support.

So, to be very explicit about how this parameter corresponds with mapping P_0 into a number, as presented in Chap. 1:

$$\Psi(P_0) = \sum_w \Big[\sum_y yP_0(Y = y \mid A = 1, W = w) - \sum_y yP_0(Y = y \mid A = 0, W = w) \Big] P_0(W = w),$$

where

$$P_0(Y = y \mid A = a, W = w) = \frac{P_0(W = w, A = a, Y = y)}{\sum_y P_0(W = w, A = a, Y = y)}$$

is the conditional probability distribution of $Y = y$, given $A = a, W = w$, and

$$P_0(W = w) = \sum_{y,a} P_0(Y = y, A = a, W = w)$$

is the marginal probability distribution of $W = w$. This statistical parameter Ψ is defined on all probability distributions of (W, A, Y). The statistical model $\mathcal{M}$ is nonparametric and $\Psi : \mathcal{M} \to \mathbb{R}$.

We note again that we use the term statistical model for the collection of possible probability distributions, while we use the word model for the statistical model augmented with the nontestable causal assumptions coded by the underlying SCM and its relation to the observed data distribution of O. In our LTPA example, the model is the nonparametric statistical model augmented with the nontestable SCM. If this model includes the randomization assumption, and the experimental treatment assignment assumption, then this model allows the identifiability of the additive causal effect $E_0Y_1 - E_0Y_0$ through the statistical target parameter $\Psi(P_0) = E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$.

2.3.5 Interpretation of Target Parameter

The observed data parameter $\Psi(P_0)$ can be interpreted in two possibly distinct ways:

1. $\Psi(P_0)$ with $P_0 \in \mathcal{M}$ augmented with the truly reliable additional nonstatistical assumptions that are known to hold (e.g., $\mathcal{M}^F$). This may involve bounding the deviation of $\Psi(P_0)$ from the desired target causal effect $\Psi^F(P_{U,X,0})$ under a realistic causal model $\mathcal{M}^F$ that is not sufficient for the identifiability of this causal effect.
2. The truly causal parameter $\Psi^F(P_{U,X}) = \Psi(P_0)$ under the more restricted SCM $\mathcal{M}^{F*}$, thereby now including all causal assumptions that are needed to make the desired causal effect identifiable from the probability distribution P_0 of O.

The purely statistical (noncausal) parameter given by interpretation 1 is often of interest, such as $E_W[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$, which can be interpreted as the average of the difference in means across the strata for W. With this parameter we can assume nothing, beyond the experimental treatment assignment assumption, except perhaps time ordering $W \rightarrow A \rightarrow Y$, to have a meaningful interpretation of the difference in means. Since we do not assume an underlying system, the SCM for (U, X) and thereby Y_a, or the randomization assumption, the parameter is a statistical parameter only. This type of parameter is sometimes referred to as a variable importance measure.

For example, if A = age, the investigator may not be willing to assume an SCM defining interventions on age (a variable one cannot intervene on and set in practice). Thus, if one does not assume $\mathcal{M}^F$, the statistical parameter $\Psi(P_0)$ under interpretation 1 can still be very much of interest. In some cases, however, these two interpretations coincide. What is known about the generation of data and distribution

P_U may imply the assumptions necessary to interpret $\Psi(P_0)$ as the causal parameter $\Psi^F(P_{U,X})$: for example, in an RCT, by design, assuming full compliance and no missingness or censoring, the causal assumptions required will hold.

2.4 Revisiting the Mortality Example

For the sake of presentation, we intentionally assumed that the exposure LTPA was binary and worked with an SCM that generated a binary exposure A. In the actual mortality study A is continuous valued. Consider the more realistic SCM $W = f_W(U_W), A = f_A(W, U_A), Y = f_Y(W, A, U_Y)$, where A is now continuous valued. Let $Y_a(u)$ be the counterfactual obtained by setting $A = a$ and $U = u$, so that Y_a is the random variable representing survival at 5 years under LTPA at level a. Suppose one wishes to consider a cut-off value δ for LTPA level so that one can recommend that the population at least exercise at this level δ. A causal quantity of interest is now

$$\psi_0^F = \sum_a w_1(a) E_0 Y_a - \sum_a w_0(a) E_0 Y_a,$$

where $w_1(a)$ is a probability distribution on excercise levels larger than δ, and $w_0(a)$ is a probability distribution on exercise levels smaller than or equal to δ. This corresponds to $E_0 Y_1 - E_0 Y_0$, where Y_1 is defined by the random intervention on the SCM in which one randomly draws A from w_1, and similarly Y_0 is defined by randomly drawing A from w_0. This causal effect $E_0 Y_1 - E_0 Y_0$ can be identified from the probability distribution P_0 of $O = (W, A, Y)$ as follows:

$$\psi_0^F = \sum_a (w_1 - w_0)(a) E_0 E_0(Y \mid A = a, W) \equiv \psi_0.$$

2.5 Road Map for Targeted Learning

In Chap. 1, we introduced the road map for targeted learning. In this chapter we have discussed defining the research question, which involved describing the data and committing to a statistical model and target parameter. The estimation problem we wish to solve is now fully defined. The next stage of the road map addresses estimation of the target parameter, which will be covered in the next three chapters.

The statistical estimation problem. We observe n i.i.d. copies $O_1, .., O_n$ from a probability distribution P_0 known to be in a statistical model $\mathcal{M}$, and we wish to infer statistically about the target parameter $\Psi(P_0)$. Often, this target parameter only depends on P_0 through a relevant (infinite-dimensional) parameter $Q_0 = Q_0(P_0)$ of P_0, so that we can also write $\Psi(Q_0)$.

Targeted substitution estimator. We construct a substitution estimator $\Psi(Q_n^*)$ obtained by plugging in an estimator Q_n^* of Q_0. This involves super learning and

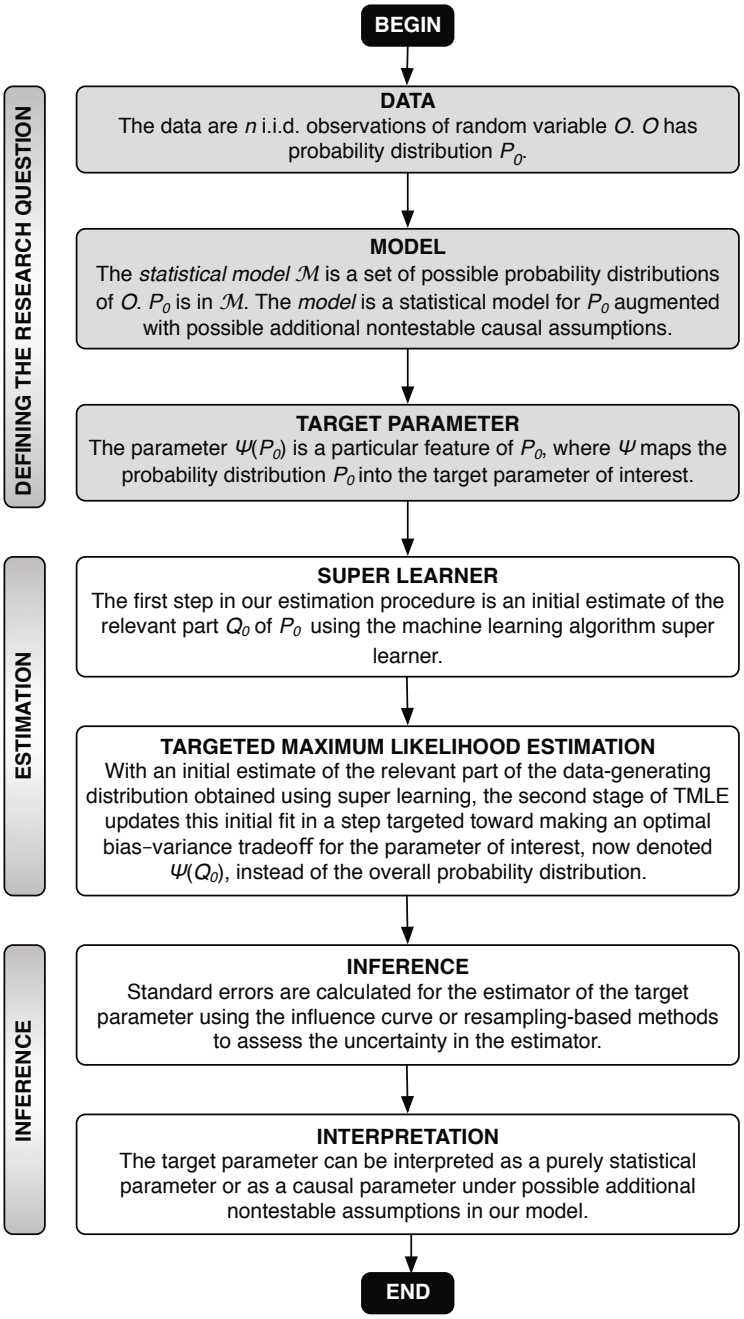

Fig. 2.6 Road map for targeted learning

TMLE, so that we obtain, under regularity conditions, an asymptotically linear, double robust, and efficient normally distributed estimator of $\psi_0 = \Psi(Q_0)$, and, in general, put in the maximal effort to minimize the mean squared error with respect to the true value ψ_0. In addition, we provide statistical inference about ψ_0 based on the estimation of the normal limit distribution of $\sqrt{n}(\Psi(Q_n^*) - \psi_0)$.

2.6 Conceptual Framework

This section provides a rigorous conceptual framework for the topics covered in this chapter. If you find it too abstract on your initial reading, we advise you to come back as you become more familiar with the material. It is meant for more advanced readers.

Data are meaningless without knowledge about the experiment that generated the data. That is, data are realizations of a random variable with a certain probability distribution on a set of possible outcomes, and statistical learning is concerned with learning something about the probability distribution of the data. Typically, we are willing to view our data as a realization of n independent identical replications of the experiment, and we accept this as our first modeling assumption. If we denote the random variable representing the data generated by the experiment by O, having a probability distribution P_0, then the data set corresponds with drawing a realization of n i.i.d. copies $O_1, \ldots, O_n$ with some common probability distribution P_0.

A statistical estimation problem corresponds with defining a statistical model $\mathcal{M}$ for P_0, where the statistical model $\mathcal{M}$ is a collection of possible probability distributions of O. The estimation problem also requires a mapping Ψ on this statistical model $\mathcal{M}$, where Ψ maps any $P \in \mathcal{M}$ into a vector of numbers $\Psi(P)$. We write $\Psi : \mathcal{M} \to \mathbb{R}^d$ for a d-dimensional parameter. We introduce ψ_0, and the interpretation of ψ_0 as $\Psi(P_0)$, i.e., a well-defined feature of P_0, is called the pure statistical interpretation of the parameter value ψ_0. The statistical estimation problem is now to map the data set $O_1, \ldots, O_n$ into an estimator of $\Psi(P_0)$ that incorporates the knowledge that $P_0 \in \mathcal{M}$, accompanied by an assessment of the uncertainty in the estimator of ψ_0.

When thinking purely about the construction of an estimator, the only concern is to construct an estimator of ψ_0 that has small mean squared error (MSE), or some other measure of dissimilarity between the estimator and the true ψ_0. This does not require any additional knowledge (or nontestable causal assumptions). As a consequence, for the construction of a targeted maximum likelihood estimator, which we introduce in Chaps. 4 and 5, the only input is the statistical model $\mathcal{M}$ and the mapping Ψ representing the target parameter.

Making assumptions about P_0 that do not change the statistical model, so-called nontestable assumptions, will not change the statistical estimation problem. However, such assumptions allows one to interpret a particular parameter $\Psi(P_0)$ in a new way. If such nontestable assumptions are known to be true, it enriches the interpretation of the number ψ_0. If they are wrong, then it results in misinterpretation of ψ_0.

This is called causal modeling when it involves nontestable assumptions that allow $\Psi(P_0)$ to be interpreted as a causal effect, and, in general, it is modeling with nontestable assumptions with the goal of providing an enriched interpretation of this parameter $\Psi(P_0)$.

It works as follows. One proposes a parameterization $\theta \rightarrow P_\theta$ for θ varying over a set Θ so that the statistical model $\mathcal{M}$ can be represented as $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, where θ represents $P_{U,X}$ in our SCM framework, but it can represent any underlying structure (not necessarily causal). That is, we provide a parameterization for the statistical model $\mathcal{M}$. In addition, since $P_0 \in \mathcal{M}$, there exists a θ_0 such that $P_0 = P_{\theta_0}$. Assume that this θ_0 is actually uniquely identified by P_0. θ_0 has its own interpretation, such as the probability distribution of counterfactual random variables in the SCM. Suddenly, the P_0 allows us to infer $\theta_0 = \Theta(P_0)$ for a mapping Θ. As a consequence, with this "magic trick" of parameterizing P_0 we succeeded in providing a new interpretation of P_0 and, in particular, of any parameter $\Psi(P_0) = \Psi(P_{\theta_0})$ as a function of θ_0.

As one can imagine, there are millions of possible magic tricks one can carry out, each one creating a new interpretation of P_0 by having it mapped into an interpretation of a θ_0 implied by a particular parameterization. The data cannot tell you if one magic trick will provide a more accurate description of reality than another magic trick, since data can only provide information about P_0 itself. As a consequence, which magic trick is applied, or if any trick is applied at all, should be driven by true knowledge about the underlying mechanism that resulted in the generation of O. In that case, the selection of the parameterization is not a magic trick but represents the incorporation of true knowledge allowing us to interpret the parameter ψ_0 for what it is. Note that this modeling could easily correspond with a nonparametric statistical model $\mathcal{M}$ for P_0.

Two important mistakes can occur in statistical practice, before the selection of an estimator, given that one has specified a statistical model $\mathcal{M}$ and parameter $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$. The first mistake is that one specifies the statistical model $\mathcal{M}$ incorrectly so that $P_0 \notin \mathcal{M}$, resulting in misinterpretation of $\Psi(P_0)$, even as a purely statistical parameter, i.e., as a mapping Ψ applied to P_0. The second mistake is that one misspecifies additional nontestable assumptions as coded by the selected parameterization for $\mathcal{M}$ that were used to provide an enriched interpretation of $\Psi(P_0)$, again resulting in misinterpretation of $\Psi(P_0)$. These two mistakes can be collapsed into one, namely, misspecification of the model for P_0. By the model we now mean the statistical model for P_0 augmented with the additional nontestable structural assumptions, even though these do not change the statistical model.

So a model now includes the additional parameterization, such that two identical statistical models that are based on different parameterizations are classified as different models. Thus, a model is defined by a mapping $P_\cdot : \Theta \rightarrow \mathcal{M}$, $\theta \rightarrow P_\theta$, and the statistical model implied by this model is given by the range $\mathcal{M} = \{P_\theta : \theta\}$ of this mapping. Regarding statistical vocabulary, we will use the word model for the parameterization mapping $P_\cdot : \Theta \rightarrow \mathcal{M}$, and statistical model for the set of possible probability distributions, i.e., the range of this mapping. Note, that if the parame-

terization is simply the identity mapping defined on $\mathcal{M}$, then the model equals the statistical model.

Even though it is healthy to be cynical about modeling and extremely aware of its dangers and its potential to lie with data, it is of fundamental importance to statistical learning that we can incorporate structural knowledge about the data-generating process and utilize that in our interpretation. In addition, even if these structural assumptions implied by the model/parameterization are uncertain, it is worthwhile to know that, *if* these were true, then our parameter would allow its corresponding interpretation. One could then report both the statistical interpretation, or the reliable statistical model interpretation, as well as the *if also, then* interpretation to our target ψ_0.

In addition, this structural modeling allows one to create truly interesting parameters in an underlying world and one can then establish under what assumptions one can identify these truly interesting parameters from the observed data. This itself teaches us how to generate new data so that these parameters will be identifiable. The identifiability results for these truly interesting parameters provide us with statistical parameters $\Psi(P_0)$ that might be interesting as statistical parameters anyway, without these additional structural assumptions, and have the additional flavor of having a particularly powerful interpretation if these additional structural assumptions happen to be true. In particular, one may be able to interpret $\Psi(P_0)$ as the best possible approximation of the wished causal quantity of interest based on the available data. Overall, this provides us with more than enough motivation to include (causal) modeling as an important component in the road map of targeted learning from data.

2.7 Notes and Further Reading

As noted in the introduction, a thorough presentation of SCMs, causal graphs, and related identifiability theory can be found in Pearl (2009). We also direct the interested reader to Judea Pearl's Web site (http://bayes.cs.ucla.edu/jp_home.html) for easily organized references and presentations on these topics. The g-formula for identifying the distribution of counterfactuals from the observed data distribution, under the sequential randomization assumption, was originally published in Robins (1986). The simplified data example we introduce in this chapter, a mortality study examining the effect of LTPA, is based on data presented in Tager et al. (1998). We carry this example through the next three chapters, and in Chap. 4, we analyze this data using super learning and targeted maximum likelihood estimation.

In our road map we utilize causal models, such as SCMs and the Neyman–Rubin model, to generate statistical effect parameters $\psi_0 = \Psi(P_0)$ of interest. The interpretation of the estimand ψ_0, beyond its pure statistical interpretation, depends on the required causal assumptions necessary for identifiability of the desired causal quantity ψ_0^F (defined as target quantity in causal model for full data or counterfactuals) from the observed data distribution. Such an interpretation might be further

enriched if one could define an actual experiment that would reproduce this causal quantity. Either way, our road map poses these causal models as working models to derive these statistical target parameters that can be interpreted as causal effects under explicitly stated causal assumptions. The latter assumptions are fully exposed and for anybody to criticize.

We wish to stress that the learning of these estimands with their pure statistical interpretation already represents progress in science. In addition, the required causal assumptions that would allow a richer interpretation of the estimand teach us how to improve our design of the observational or RCT.

Somehow, we think that a statistical target parameter that has a desired causal interpretation under possibly unrealistic assumptions is a "best" approximation of the ideal causal quantity, given the limitations set by the available data. For example, $E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$ is an effect of treatment, controlling for the measured covariates, with a clear statistical interpretation, and, if people feel comfortable talking about $E_0Y_1 - E_0Y_0$, then we think that this statistical estimand represents a "best" effort to target this additive causal effect under the constraints set by the available data.

Instead of making a hard decision regarding the causal assumptions necessary for making the estimand equal to the causal quantity, one may wish to investigate the potential distance between the estimand and the causal quantity. In this manner, one still allows for a causal interpretation of the estimand (such as that the asymptotic bias of the estimand with respect to the desired causal quantity is bounded from above by a certain number), even if the causal assumptions required for making the estimand equal to the causal quantity are violated. Such an approach relies on the ability to bound this distance by incorporation of realistic causal knowledge. Such a sensitivity analysis will require input from subject matter people such as a determination of an upper bound of the effect of unmeasured confounders beyond the measured time-dependent confounders. Even a highly trained statistician will have an extremely hard time getting his/her head around such a question, making such sensitivity analyses potentially unreliable and extremely hard to communicate. Still, this is an important research area since it allows for a continuous range from pure statistical interpretation of the estimand to a pure causal effect interpretation.

Either way, we should not forget that using poor methods for estimation with the actual observed data, while investing enormous effort in such a sensitivity analysis. makes no sense. By the same token, estimation of the estimand is a separate problem from determining the distance between the estimand and the causal quantity of interest and is obviously as important as carefully defining and interpreting the estimand: the careful definition and interpretation of an estimand has little value if one decides to use a misspecified parametric model to fit it!

http://www.springer.com/978-1-4419-9781-4

Targeted Learning
Causal Inference for Observational and Experimental Data
van der Laan, M.J.; Rose, S.
2011, LXXII, 628 p., Hardcover
ISBN: 978-1-4419-9781-4