Journal de la Société Française de Statistique
Vol. 161 No. 1 (2020)

A Primer on Causality in Data Science

Hachem Saddiki and Laura B. Balzer

Abstract: Many questions in Data Science are fundamentally causal in that our objective is to learn the effect of some exposure, randomized or not, on an outcome interest. Even studies that are seemingly non-causal, such as those with the goal of prediction or prevalence estimation, have causal elements, including differential censoring or measurement. As a result, we, as Data Scientists, need to consider the underlying causal mechanisms that gave rise to the data, rather than simply the pattern or association observed in those data. In this work, we review the "Causal Roadmap" of Petersen and van der Laan (2014) to provide an introduction to some key concepts in causal inference. Similar to other causal frameworks, the steps of the Roadmap include clearly stating the scientific question, defining of the causal model, translating the scientific question into a causal parameter, assessing the assumptions needed to express the causal parameter as a statistical estimand, implementation of statistical estimators including parametric and semi-parametric methods, and interpretation of our findings. We believe that using such a framework in Data Science will help to ensure that our statistical analyses are guided by the scientific question driving our research, while avoiding over-interpreting our results. We focus on the effect of an exposure occurring at a single time point and highlight the use of targeted maximum likelihood estimation (TMLE) with Super Learner.

Keywords: Causal inference, Directed acyclic graphs (DAGs), Observational studies, Structural causal models, Targeted learning, Targeted maximum likelihood estimation (TMLE)
AMS 2000 subject classifications: 62-01, 62-07, 62A01, 62P10

1. Introduction

Recently, Hernán et al. (2018) classified Data Science into three tasks: description, prediction, and causal inference. The first two fall firmly in the realm of statistical inference in that they are purely data-driven tasks, while the last requires something more than the observed data alone (Pearl et al., 2016). Consider, for example, the target population of HIV-infected women of child-bearing age (15-49 years old) in East Africa. After obtaining measurements on a sample of women from this population, we could provide some basic descriptive statistics on demographic and clinical variables, such as age, education, use of antiretroviral therapy, pregnancy, and viral suppression, defined as plasma HIV RNA <500 copies/mL. Likewise, we could use these variables to build a predictor of viral suppression. This predictor could rely on parametric logistic regression or more advanced machine learning algorithms, such as Super Learner (van der Laan et al., 2007; Petersen et al., 2015).

Now consider the potential impact of pregnancy on clinical outcomes in this population. While optimizing virologic outcomes is essential to preventing mother-to-child-transmission of HIV, the prenatal period could plausibly disrupt or enhance HIV care for a pregnant woman (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2014). We can then ask, *"what is the effect of pregnancy on HIV RNA viral suppression among HIV-positive women of child-bearing*

Department of Biostatistics & Epidemiology, University of Massachusetts-Amherst, 715 North Pleasant St. Amherst, MA 01003-9304.
E-mail: hsaddiki@umass.edu and E-mail: lbalzer@umass.edu

age in East Africa?". While the exposure of pregnancy is not a traditional treatment as commonly considered in a randomized trial, this question is still causal in that we are asking about the outcomes of patients under two different conditions and to answer this question, we must go beyond the observed data set.

In particular, causal inference requires an *a-priori* specified set of, often untestable, assumptions about the data generating mechanism. Once we posit a causal model, often encoded in the language of causal graphs, we can express our scientific question in terms of a causal quantity. Under explicit assumptions, we can then translate that causal quantity into a statistical estimand, a function of the observed data distribution. This translation, called *identifiability*, is not guaranteed, as it depends on the underlying scientific question, the structure of the causal model, and the observed data. Lack of identifiability, however, provides us guidance on further data collection efforts and the additional assumptions needed for such translation. Altogether we obtain a statistical estimand that as closely as possible matches the underlying scientific question and thereby ensures that our objective is driving the statistical analysis (Petersen and van der Laan, 2014; Hernán et al., 2008). Once the estimand has been specified, we return to realm of statistics and the purely data-driven exercises of point estimation, hypothesis testing, and creating confidence intervals. Interpretation of the resulting values, however, requires us again to consider our causal assumptions.

In this primer, we review the Causal Roadmap of Petersen and van der Laan (2014) to (1) specify the scientific question; (2) build an accurate causal model of our knowledge; (3) define the target causal quantity; (4) link the observed data to the causal model; (5) assess identifiability; (6) estimate the resulting statistical parameter; and (7) appropriately interpret the results. This Roadmap borrows the general logic from Descartes's Scientific Method (Descartes, 1637) and shares a common flow of other causal frameworks (Neyman, 1923; Rubin, 1974; Holland, 1986; Robins, 1986; Rubin, 1990; Spirtes et al., 1993; Pearl, 2000; Little and Rubin, 2000; Dawid, 2000; Heckman and Vytlacil, 2007; Robins and Hernán, 2009; van der Laan and Rose, 2011; Richardson and Robins, 2013; Hernán and Robins, 2016). In particular, all approaches demand a clear statement of the research objective, including the target population and interventions of interest (Hernán, 2018; Ahern, 2018). All approaches also provide guidance for conducting a statistical analysis that best answers the motivating question. Unlike some of the other frameworks, however, the Roadmap emphasizes the use of non-parametric or semi-parametric statistical methods, such as targeted maximum likelihood estimation described in Section 2.6, to avoid unwarranted parametric assumptions and harness recent advances in machine learning. As a result this framework has sometimes been called the Targeted Learning Roadmap (van der Laan and Rose, 2011; Tran et al., 2016; Kreif et al., 2017).

2. The Roadmap for Causal Inference

2.1. Specify the Scientific Question

The first step is to specify our scientific question. This helps frame our objective in a more detailed way, while incorporating knowledge about the study. In particular, we need to specify the target population, the exposure, and the outcome of interest. As our running example, we ask, what is the effect of becoming pregnant on HIV RNA viral suppression (<500 copies/mL) among HIV-positive women of child-bearing age (15-49 years) in East Africa?

This question provides a clear definition of the study variables and objective of our research. It also makes explicit that the study only makes claims about the effect of a specific exposure, outcome, and target population. Any claims outside this context, such as a different exposure, outcome, or target population, represent distinct questions and would require going through the Roadmap again from the start. The temporal cues present in the research question are of particular importance. They represent the putative cause, here pregnancy, and effect of interest, here viral suppression. The temporal cues, together with background knowledge, are frequently used as a basis for specifying the causal model, our next step.

2.2. *Specify the Causal Model*

One of the appealing features of causal modeling, and perhaps the reason behind its success, is the rich and flexible language for encoding mechanisms underlying a data generating process. Here, we focus on Pearl (2000)'s structural causal models, which unify causal graphs and structural equations (Pearl, 1988; Goldberger, 1972; Duncan, 1975). Structural causal models formalize our knowledge, however limited, of the study, including the relationships between variables and the role of unmeasured factors.

Let us consider again our running example of the impact of pregnancy on HIV viral suppression among women in East Africa. Let W_1 denote the set of baseline demographic covariates, such as age, marital status, and education level, and W_2 denote the set of pre-exposure HIV care variables, such as prior use of antiretroviral therapy. The exposure A is a binary variable indicating that the woman is known to be pregnant, and the outcome Y is a binary indicator of currently suppressing HIV viral replication: <500 copies/mL. These constitute the set of *endogenous* variables, which are denoted $X = \{W_1, W_2, A, Y\}$ and are essential to answering the research question.

Each endogenous variable is associated with a latent background factor U_{W_1}, U_{W_2}, U_A, and U_Y, respectively. The set of background factors are called *exogenous* variables and denoted $U = (U_{W_1}, U_{W_2}, U_A, U_Y)$. These variables account for all other unobserved sources that might influence each of the endogenous variables and can share common components. In our example, unmeasured background factors U might include socioeconomic status, the date of HIV infection, the date of conception, her partner's HIV status, and her genetic profile.

Causal Graphs: The "causal story" of the data can be conveyed using the language of graphs (Pearl, 2000; Pearl et al., 2016). Graphical models consist of a set of nodes representing the variables, and a set of directed or undirected edges connecting these nodes. Two nodes are *adjacent* if there exists an edge between them, and a *path* between two nodes A and B is a sequence of adjacent nodes starting from A and ending in B. If an edge is *directed* from node A to node B, then A is the *parent* of B, and B is the *child* of A. More generally, for any path starting from node A, the set of nodes included in this path are *descendants* of A, and A is the *ancestor* of all the nodes included in this set.

Here, we are interested in Directed Acyclic Graphs (DAGs), which are fully directed graphs with no path from a given node to itself. DAGs provide a mechanism to explicitly encode our causal assumptions about the underlying data generating process. Specifically, a variable A is a *direct cause* of another variable B, if B is the child of A in the causal graph. Also, a variable A is a *cause* of another variable B, if B is a descendant of A in the causal graph (Pearl, 2000).

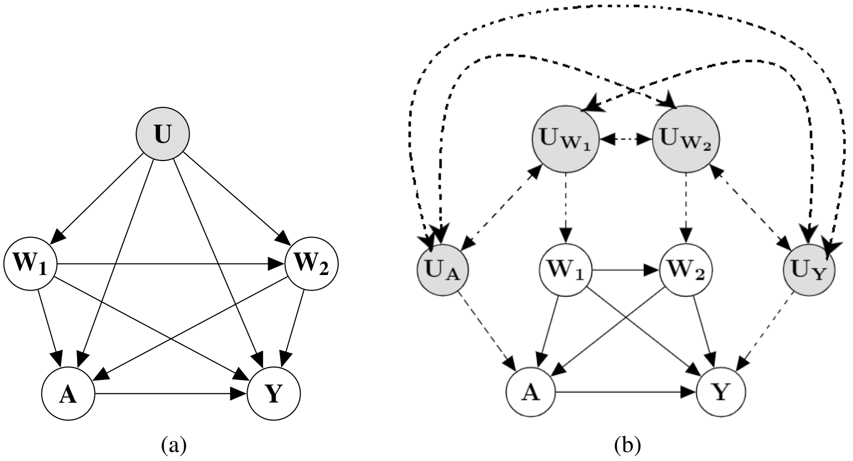

Figure 1: Encoding the underlying causal mechanisms with graphical models. Shaded nodes represent exogenous variables, and unshaded nodes are endogenous variables. Directed edges represent a direct cause between a pair of variables. Double-headed dashed arrows represent potential correlation between the exogenous factors (i.e., unmeasured common causes of the endogenous variables). In (a) we give a directed acyclic graph (DAG) with a single node U representing all the common unmeasured sources. In (b) we provide an alternative representation to make explicit the relationships between the unmeasured background factors $U = \{U_{W_1}, U_{W_2}, U_A, U_Y\}$ and each endogenous variable.

To illustrate, Figure 1(a) provides a DAG corresponding to our running example. From this graph, we can make the following statements:

1. Baseline demographics W_1 may affect a woman's pre-exposure HIV care W_2, her pregnancy status A, and her HIV viral suppression status Y.
2. Prior care W_2 may affect her pregnancy status A, and her HIV viral suppression status Y.
3. Being pregnant A may affect her HIV viral suppression status Y.
4. Unmeasured factors $U = (U_{W_1}, U_{W_2}, U_A, U_Y)$ may affect a woman's baseline characteristics, her prior care, her fertility, and her suppression outcome.

In Figure 1(a), a single node U represents all the common, unmeasured factors that could impact the pre-exposure covariates, the exposure, and the outcome. In an alternative representation in Figure 1(b), we have explicitly shown each exogenous variable $(U_{W_1}, U_{W_2}, U_A, U_Y)$ as a separate node and as parent to its corresponding endogenous variable (W_1, W_2, A, Y), respectively. In the latter, dashed double-headed arrows denote correlation between the exogenous factors.

Both representations make explicit that there could be unmeasured common causes of the covariates $W = (W_1, W_2)$ and the exposure A, the exposure A and the outcome Y, and the covariates W and the outcome Y. In other words, there is measured and unmeasured confounding present in this study. Altogether, we have avoided many unsubstantiated assumptions about the causal relationships between the variables. This causal model is, thus, non-parametric beyond the assumed time-ordering between variables.

Causal graphs can be extended to accommodate more complicated data structures. Suppose, for example, plasma HIV RNA viral levels are missing for some women in our population of interest. We could modify our causal model to account for incomplete measurement (Robins et al., 2000, 1994; Scharfstein et al., 1999; Daniel et al., 2012; Mohan et al., 2013; Balzer et al., 2017). Specifically, we redefine the exposure node for pregnancy as A_1 and introduce a new intervention node A_2 defined as indicator that her viral load is measured. The resulting causal graph is represented in Figure 2. We refer the readers to Mohan et al. (2013) for detailed discussion of formulating a causal model for the missingness mechanism and to Petersen et al. (2017) for a real world application handling missingness on both HIV status and viral loads (Balzer et al., 2017). For the remainder of the primer, we assume, for simplicity, there are no missing data and Figure 1 holds. As discussed in the Appendix, other extensions can also be made to account for common complexities, such as longitudinal data and effect mediation.

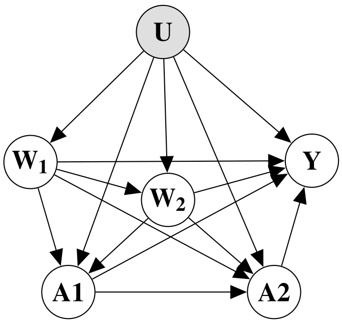

Figure 2: Causal graph extending the running example to account for missingness on the outcome. Along the baseline demographics W_1, clinical covariates W_2, and suppression outcome Y, we now have two intervention nodes A_1 for pregnancy status and A_2 for measurement of plasma HIV RNA level.

In the subsequent steps, we discuss how altering the causal graph, particularly by removing edges, is equivalent to making additional assumptions about the data generating process. Before doing so, however, we present the causal model in its structural form.

Non-Parametric Structural Equations: Structural causal models also encode information about the data generating process with a set of non-parametric equations. Like the causal graph, these equations describe how “nature” would deterministically generate the variables in our study (Pearl, 2000; Pearl et al., 2016). Use of the equations can be preferable in longitudinal settings when causal graphs can become unwieldily.

Formally, we define a structural causal model, denoted $\mathscr{M}^*$, by the set of exogenous variables U, the set of endogenous variables X, and a set of functions $\mathscr{F}$ that deterministically assign a value to each variable in X, given as input the values of other variables in X and U. These non-parametric structural equations allow us to expand our definition of causal assumptions (Pearl, 2000; Pearl et al., 2016). Variable A is considered to be a *direct cause* of variable B, if A appears in the function assigning a value to B. Variable A is also a *cause* of variable B, if A is direct cause of B or any causes of B.

In our HIV viral suppression example, the corresponding structural equations are

$$\begin{aligned} W_1 &= f_{W_1}(U_{W_1}) \\ W_2 &= f_{W_2}(W_1, U_{W_2}) \\ A &= f_A(W_1, W_2, U_A) \\ Y &= f_Y(W_1, W_2, A, U_Y) \end{aligned} \tag{1}$$

where the set of functions $\mathscr{F} = \{f_{W_1}, f_{W_2}, f_A, f_Y\}$ encode the mechanism deterministically generating the value of each endogenous variable. The exogenous variables $U = \{U_{W_1}, U_{W_2}, U_A, U_Y\}$ have a joint probability distribution $\mathbb{P}_U$ and coupled with the set of structural equations $\mathscr{F}$ give rise to a particular data generating process that is compatible with the causal assumptions implied by $\mathscr{M}^*$.

In our example, for a given probability distribution $\mathbb{P}_U$ and set of structural equations $\mathscr{F}$, the structural causal model $\mathscr{M}^*$ describes the following data generating process. For each woman,

1. *Draw the exogenous variables U from the joint probability distribution $\mathbb{P}_U$.* Intuitively, when we sample a woman from the population, we obtain all the unmeasured variables that could influence her baseline covariates, prior care, pregnancy status, and suppression outcome.
2. *Generate demographic covariates W_1 deterministically using U_{W_1} as input to the function f_{W_1}*; the demographic covariates include her age, marital status, education attained, and socioeconomic status.
3. *Generate past HIV care covariates W_2 deterministically using U_{W_2} and the woman's demographic covariates W_1 as input to the function f_{W_2}*; the measured clinical factors include history of antiretroviral therapy use and prior HIV suppression status.
4. *Generate pregnancy status A deterministically using U_A, W_1, and W_2 as inputs to function f_A.* Recall A is an indicator equaling 1 if the woman is known to be pregnant and 0 otherwise.
5. *Generate HIV suppression outcome Y deterministically using U_Y, W_1, W_2, and A as inputs to function f_Y.* Recall Y is an indicator equaling 1 if her HIV RNA viral level is less than 500 copies/mL and 0 otherwise.

It is important to note that the set of structural equations are non-parametric. In other words, the explicit relationship between the system variables, as captured by the set of functions $\mathscr{F}$, are left unspecified. If knowledge is available regarding a relationship of interest, it can be readily incorporated in the structural equations. For instance, in a two-armed randomized trial with equal allocation probability, the function that assigns a value to the exposure variable A can be explicitly encoded as $A = f_A(U_A) = \mathbb{I}(U_A < 0.5)$, where $\mathbb{I}$ is an indicator function and U_A assumed to be drawn from a $Uniform(0,1)$.

2.3. Define the Target Causal Quantity

Once the causal model is specified, we may begin to ask questions of causal nature. The rationale comes from the observation that the structural causal model $\mathscr{M}^*$ is not restricted to the particular setting of our study, but can also describe the same system under changed conditions. The structural equations are *autonomous*, which means that modifying one function does not change another. Therefore, we can make targeted modifications to our causal model to evaluate hypothetical, counterfactual scenarios that would otherwise never be realized, but correspond to our underlying scientific question.

In our running example, we are interested in the effect of pregnancy on viral suppression. In the original causal model (Figure 1 and Equation 1), a woman's pregnancy status is determined by her baseline demographics W_1, prior care status W_2, and unmeasured factors U_A, such as contraceptive use. However, our objective is to determine the probability of viral suppression if all women in the target population were pregnant versus if the same women over the same time-frame were not pregnant. The autonomy of the structural equations allows us to modify the way in which the exposure, here pregnancy, is determined. In particular, we can intervene on the exposure A to deterministically set $A = 1$ in one scenario, and then set $A = 0$ in another, while keeping the other equations constant.

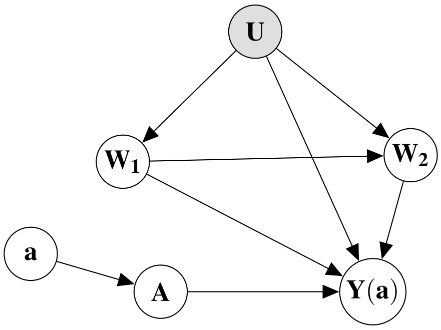

Figure 3: Causal graph after intervention on the exposure, pregnancy status, to set $A = a$. Since this is done deterministically and independently of other variables in the system, the only node *causing* a change in A is the intervention node $a \in \{0,1\}$.

The post-intervention causal graph is given in Figure 3 and the structural equations become

$$\begin{aligned} W_1 &= f_{W_1}(U_{W_1}) & \qquad W_1 &= f_{W_1}(U_{W_1}) \\ W_2 &= f_{W_2}(W_1, U_{W_2}) & \qquad W_2 &= f_{W_2}(W_1, U_{W_2}) \\ A &= 1 & \qquad A &= 0 \\ Y(1) &= f_Y(W_1, W_2, 1, U_Y) & \qquad Y(0) &= f_Y(W_1, W_2, 0, U_Y) \end{aligned}$$

These interventions generate counterfactual outcomes $Y(a)$ for $a \in \{0,1\}$, whose distribution is denoted $\mathbb{P}^*$. These causal quantities are indicators that a participant would have suppressed viral replication, if possibly contrary to fact, her pregnancy status were $A = a$.

In this case, it is both physically impossible and unethical to design a randomized trial for pregnancy. In other words, we cannot directly intervene on a woman's pregnancy status. Likewise in Figure 2, enforcing measurement of the outcome, which translates into setting $A_2 = 1$, is impossible. While neither intervention is plausible, we believe counterfactuals provide a language to express many questions in Data Science in a mathematically tractable way. Nonetheless, we note that there has been a lively debate about defining and interpreting analyses of variables on which one cannot directly intervene (Pearl, 1995; Hernán, 2005; van der Laan et al., 2005; Hernán and VanderWeele, 2011; Petersen, 2011; Petersen and van der Laan, 2014; Hernán and Robins, 2016).

Given the counterfactual outcomes and their distribution $\mathbb{P}^*$, we can express our scientific question as a mathematical quantity. One common choice is the Average Treatment Effect (ATE):

$$\Psi^*(\mathbb{P}^*) := \mathbb{E}^*[Y(1) - Y(0)], \tag{2}$$

where the expectation is taken with respect to $\mathbb{P}^*$. Since the causal model $\mathscr{M}^*$ provides the set of possible probability distributions for the exogenous and endogenous factors (U,X) and thus the counterfactual outcomes $(Y(1),Y(0))$, Ψ^* is a mapping from $\mathscr{M}^*$ to the real numbers. The target causal parameter $\Psi^*(\mathbb{P}^*)$ represents the difference in the expected counterfactual outcome if all units in the target population were exposed and the expected counterfactual outcome if the same units were not exposed. For the running example, $\Psi^*(\mathbb{P}^*)$ can be interpreted as the difference in the counterfactual probability of viral suppression if all women in the target population were pregnant versus if the same women were not.

Before discussing how these causal quantities can be identified from the observed data distribution, we emphasize that for simplicity we have focused on a binary intervention, occurring deterministically at a single time point. Scientific questions corresponding to categorical, continuous, stochastic, and longitudinal exposures are also encompassed in this framework, but beyond the scope of this primer and are briefly discussed in the Appendix. We also note that other summaries, such as relative measures, the sample average effect, or marginal structural models, may better capture the researcher's scientific question.

2.4. Link the Observed Data to the Causal Model

Thus far, we have defined our scientific question, specified a structural causal model $\mathscr{M}^*$ to represent our knowledge of the data generating process, intervened on that causal model to generate counterfactual outcomes, and used these counterfactuals to express our scientific question as a causal quantity. The next step is to provide an explicit link between the observed data and the specified structural causal model.

Returning to our running example, suppose we have a simple random sample of N women from our target population. On each woman, we measure her baseline demographics W_1, prior HIV care W_2, pregnancy status A, and suppression outcome Y. These measurements constitute our observed data for each woman in our sample: $O = \{W_1, W_2, A, Y\}$. Therefore, we have N independent, identically distributed copies of O, which are drawn from some probability distribution $\mathbb{P}$. Other sampling schemes, such as case-control, are accommodated by this framework, but are beyond the scope of this primer.

If we believe that our causal model accurately describes the data generating process, we can assume that the observed data are generated by sampling repeatedly from a distribution compatible with the structural causal model. In other words, the structural causal model $\mathscr{M}^*$ provides a description of the study under existing conditions (i.e., the real world) and under specific intervention (i.e., the counterfactual world). As a result, the observed outcome Y equals the counterfactual outcome $Y(a)$ when the observed exposure A equals the exposure of interest, a; this is commonly called the *consistency assumption*.

In our example, all the endogenous variables are observed: $X = O$; therefore, we can write

$$\mathbb{P}(O=o) = \sum_u \mathbb{P}^*(X=x|U=u)\mathbb{P}^*(U=u), \tag{3}$$

where an integral replaces the summation for continuous variables. This, however, might not always be the case. Suppose, for example, we only measured demographics, pregnancy status, and viral suppression, but failed to measure variables related to prior HIV care. Then the observed data would be $O = (W_1, A, Y)$ and are a subset of all the endogenous variables X. In either case, we see that the structural causal model $\mathscr{M}^*$, defined as the collection of all possible joint distributions of the exogenous and endogenous variables (U,X), implies the statistical model $\mathscr{M}$, defined as the collection of all possible joint distributions for the observed data O. The structural causal model $\mathscr{M}^*$ rarely implies restrictions on the resulting statistical model $\mathscr{M}$, which is thereby often non-parametric. An important exception is a completely randomized trial, where the unmeasured factors determining the treatment assignment U_A are independent of the others and results in a semi-parametric statistical model. The *D-separation* criteria of Pearl (2000) can be used to evaluate what statistical assumptions, if any, are implied by the causal model. The true observed data distribution $\mathbb{P}$ is an element of the statistical model $\mathscr{M}$.

2.5. Assessing Identifiability

In the previous section, we established a bridge between our structural causal model $\mathscr{M}^*$ and our statistical model $\mathscr{M}$. However, we have not yet discussed the conditions under which causal assumptions and observed data can be combined to answer causal questions. Structural causal models provide one way to assess the assumptions needed to express our target causal quantity as a statistical estimand, which is a well-defined function of the observed data distribution $\mathbb{P}$.

Recall in Section 2.3 that we defined our target causal parameter as the average treatment effect $\Psi^*(\mathbb{P}^*) = \mathbb{E}^*[Y(1) - Y(0)]$: the difference in the expected viral suppression status if all women were pregnant versus if none were. If given a causal model and its link to the observed data, the target causal parameter can be expressed as a function of the observed data distribution $\mathbb{P}$, then the causal parameter is called $identifiable$. If not, we can still explicitly state and evaluate the assumptions needed to render the target causal parameter identifiable from the observed data distribution.

One of the main tools for assessing identifiability of causal quantities is a set of criteria based on causal graphs. In general, these criteria provide a systematic approach to identify an appropriate adjustment set. Here, we focus on identifiability for the effect of a single intervention at one time, sometimes called “point-treatment effects”. For these problems, we first present the back-door criterion and the front-door criterion. For a detailed presentation of graphical methods for assessing identifiability in causal graphs, the reader is referred to Pearl (2000); Pearl et al. (2016).

Formally, we say that a path is *blocked* if at least one variable in that path is conditioned on, and we define a *back-door path* from a given node A as any path that contains an arrow into node A. Then, given any pair of variables (A,B), where A occurs before B in a directed acyclic graph, a set of variables C is said to satisfy the *back-door criterion* with respect to (A,B) if (1) the descendants of A do not include any node in C, and (2) C blocks every back-door path from A to B . The rationale behind this criterion is that, for C to be the appropriate adjustment set that isolates the causal effect of A on B, we must block all spurious paths between A and B, and leave directed paths from A to B unblocked. This criterion does not, however, cover all possible graph structures.

Alternatively, a set of variables C satisfies the *front-door criterion* with respect to a pair of variables (A,B) if (1) all directed paths from A to B are blocked by C, (2) all paths from A to C are blocked, and (3) all paths from C to B containing an arrow into C are blocked by A. We note that the front-door criterion is more involved than its back-door counterpart, in the sense that it requires more stringent conditions to hold for a given adjustment set to satisfy identifiability. In practice, it is often the case that the back-door criterion is enough to identify the needed adjustment set, especially in point-treatment settings. When the back-door criterion holds, the observed association between the exposure and outcome can be attributed to the causal effect of interest, as opposed to spurious sources of correlation.

In our running example, the set of baseline covariates $W = (W_1, W_2)$ will satisfy the back-door criterion with respect to the effect of pregnancy A on HIV viral suppression Y, if the following two conditions hold:

1. No node in W is a descendant of A.
2. All back-door paths from A to Y are blocked by W.

Looking at the posited causal graph from Figure 1(a), we see that the first condition holds, but the second is violated. There exists a back-door path from A to Y through the unmeasured background factors U. Intuitively, the unmeasured common causes of pregnancy and HIV viral suppression obstruct our isolation of the causal effect of interest and thus "confound" our analyses. Therefore, our target causal quantity is not identifiable in the original causal model $\mathscr{M}^*$.

Nonetheless, we can explicitly state and consider the plausibility of the causal assumptions needed for identifiability. In particular, the following independence assumptions are sufficient to satisfy the back-door criterion and thus identify the causal effect in this point-treatment setting.

1. There must not be any unmeasured common causes of the exposure and the outcome: $U_A \perp\!\!\!\perp U_Y$ *and*,
   (a) There must not be any unmeasured common causes of the exposure and the baseline covariates: $U_A \perp\!\!\!\perp U_{W_1}$ and $U_A \perp\!\!\!\perp U_{W_2}$
   or
   (b) There must not be any unmeasured common causes of the baseline covariates and the outcome: $U_{W_1} \perp\!\!\!\perp U_Y$ and $U_{W_2} \perp\!\!\!\perp U_Y$.

These criteria are reflected in the causal graphs shown in Figure 4. In the running example, assumption 1.a states that there are no unmeasured common causes of pregnancy status and demographic or clinical factors, while 1.b assumes that there are no unmeasured common causes of viral suppression and demographic or clinical factors.

The independence assumptions in 1.a hold by design in a stratified, randomized trial, where the unmeasured factors determining the exposure assignment are independent of all other unmeasured factors. As a result, these independence assumptions (1.a and/or 1.b) are sometimes called the *randomization assumption* and equivalently expressed as $Y(a) \perp\!\!\!\perp A \mid W$. These assumptions are also referred to as “unconfoundedness”, “selection on observables”, and “conditional exchangeability” (Robins, 1986).

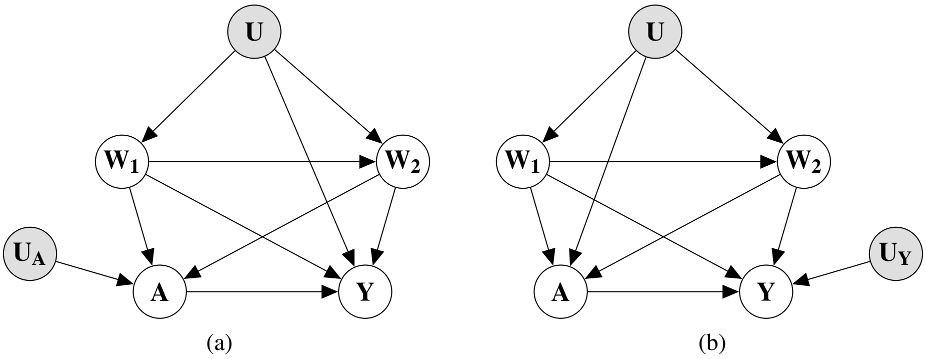

Figure 4: Causal graphs (a) and (b) encode identifiability assumptions (1.a) and (1.b), respectively. Here, we have explicitly shown that the unmeasured factors contributing to the exposure U_A in (a) and the outcome U_Y in (b) are independent of the others.

With these assumptions, we can express the distribution of counterfactual outcomes in terms of the distribution of the observed data:

$$\begin{aligned}\mathbb{P}^*(Y(a)) &= \sum_w \mathbb{P}^*(Y(a)|W=w)\mathbb{P}^*(W=w)\\ &= \sum_w \mathbb{P}^*(Y(a)|A=a,W=w)\mathbb{P}^*(W=w)\\ &= \sum_w \mathbb{P}(Y|A=a,W=w)\mathbb{P}(W=w)\end{aligned}$$

where $W=(W_1,W_2)$ denotes the pre-exposure covariates, including both demographic and clinical factors, and where the summation generalizes to an integral for continuous covariates here and in all subsequent expressions. The first equality is by the law of iterated expectations. The second equality holds by the randomization assumption, and the final by the established link between the causal and statistical model (Section 2.4).

Under these assumptions, we can express the average treatment effect $\Psi^*(\mathbb{P}^*) = \mathbb{E}^*[Y(1) - Y(0)]$, as a statistical estimand, often called the *G-computation identifiability result* (Robins, 1986):

$$\Psi(\mathbb{P}) := \sum_w \big[\mathbb{E}(Y|A=1,W=w) - \mathbb{E}(Y|A=0,W=w)\big]\mathbb{P}(W=w) \tag{4}$$

Thus, our statistical target is the difference in the expected outcome, given the exposure and covariates, and the expected outcome, given no exposure and covariates, averaged with respect to the distribution of the baseline covariates W. In our example, $\Psi(\mathbb{P})$ is the difference in the probability of viral suppression, between pregnant and non-pregnant women with the same values of the covariates, standardized with respect to the covariate distribution in the population.

The same quantity can be expressed in inverse probability weighting form:

$$\Psi(\mathbb{P}) := \mathbb{E}\left[\left(\frac{\mathbb{I}(A=1)}{\mathbb{P}(A=1\mid W)} - \frac{\mathbb{I}(A=0)}{\mathbb{P}(A=0\mid W)}\right)Y\right] \tag{5}$$

The latter representation highlights an additional data support condition, known as *positivity*:

$$min_{a\in\mathscr{A}}\ \mathbb{P}(A=a|W=w)>0,\ \text{for all } w \text{ such that } \mathbb{P}(W=w)>0.$$

Each exposure level of interest must occur with a positive probability within the strata of the discrete-valued adjustment set W. This assumption is also called "overlap" and "the experimental treatment assignment assumption". We refer the reader to Petersen et al. (2012), Cole and Hernán (2008) and Messer et al. (2010) for a discussion of this assumption and approaches when it is theoretically or practically violated.

Overall, the identifiability step is essential to specifying the needed adjustment set, and thereby statistical estimand to link our causal effect of interest to some function of the observed data distribution. Above, we focused on a simple point-treatment setting with measured and unmeasured confounding, but without mediation, biased sampling, or missing data. In more realistic settings, there are many other sources of association between our exposure and outcome, including selection bias, direct and indirect effects, and the common statistical paradoxes of Berkson's bias and Simpson's Paradox (Hernán et al., 2004; Hernández-Díaz et al., 206). Furthermore, in the setting of longitudinal exposures with time-dependent confounding, the needed adjustment set may not be intuitive and the short-comings of traditional approaches become more pronounced (Robins, 1986; Robins et al., 2000; Robins and Hernán, 2009; Pearl et al., 2016). Indeed, methods to distinguish between correlation and causation are crucial in the era of "Big Data", where the number of variables is growing with increasing volume, variety, and velocity (Rose, 2012; Marcus and Davis, 2014; Balzer et al., 2016).

Nonetheless, it is important to note that specifying a causal model (Section 2.2) does not guarantee the identification of a causal effect. Causal frameworks do, however, provide insight into the limitations and full extent of the questions that can be answered given the data at hand. They further facilitate the discussion of modifications to the study design, the measurement additional variables, and sensitivity analyses (Robins et al., 1999; Imai et al., 2010; VanderWeele and Arah, 2011; Díaz and van der Laan, 2013b).

In fact, even if the causal effect is not identifiable (e.g., Figure 1), the Causal Roadmap still provides us with a statistical estimand (e.g., Equation 4) that comes as close as possible to the causal effect of interest given the limitations in the observed dataset. In the next sections, we discuss estimation of this statistical parameter and use identifiability results, or lack there of, to inform the strength of our interpretations.

2.6. Estimate the Target Statistical Parameters

Once the statistical model and estimand have been defined, the Causal Roadmap returns to traditional statistical inference to estimate functions of a given observed data distribution. Here, we focus on estimators based on the G-computation identifiability result $\Psi(\mathbb{P})$. Popular methods for estimation and inference for $\Psi(\mathbb{P})$, which would equal the average treatment effect if the identifiability assumptions held, include parametric G-computation, Inverse Probability Weighting (IPW), and Targeted Maximum Likelihood Estimation (TMLE) (Robins, 1986; Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983; van der Laan and Rubin, 2006; van der Laan and Rose, 2011). Below we briefly outline the implementation of each estimator and refer the reader to Petersen and Balzer (2014) for worked R code for each algorithm. We emphasize that while each algorithm is targeting a causally motivated statistical estimand, these algorithms are not directly estimating causal effects, and therefore it is a misnomer to call them "causal estimators".

Parametric G-computation is an algorithm that simply estimates the quantities needed to calculate the statistical estimand defined in Equation (4) and then substitutes those quantities into the G-computation formula (Robins, 1986; Taubman et al., 2009; Young et al., 2011; Westreich et al., 2012; Zhang et al., 2018). As a result, this algorithm is sometimes called the *simple substitution estimator* and is implemented with the following steps.

1. Regress the outcome on the exposure and covariate adjustment set to estimate the conditional expectation $\mathbb{E}(Y|A,W)$.
2. Based on the estimates from Step 1, generate the predicted outcomes for each individual in the sample while deterministically setting the value of the exposure to the levels of interest, but keeping the covariates the same:

   $$\hat{\mathbb{E}}(Y_i|A_i=1,W_i) \text{ and } \hat{\mathbb{E}}(Y_i|A_i=0,W_i) \text{ for all observations } i=1,...,N.$$

   For a binary outcome, this step corresponds to generating the predicted probabilities $\hat{\mathbb{P}}(Y_i = 1|A_i = a, W_i)$ for exposure levels $a \in \{0,1\}$.
3. Obtain a point estimate by taking a sample average of the difference in the predicted outcomes from Step 2:

   $$\hat{\Psi}_{Gcomp}(\hat{\mathbb{P}}) = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mathbb{E}}(Y_i|A_i=1,W_i) - \hat{\mathbb{E}}(Y_i|A_i=0,W_i)\right]$$

   where $\hat{\mathbb{P}}$ denotes the empirical distribution, and the sample proportion is a simple non-parametric estimator of the covariate distribution $\mathbb{P}(W)$: $\hat{\mathbb{P}}(W=w) = \frac{1}{N}\sum_i \mathbb{I}(W_i = w)$.

Inverse Probability Weighting (IPW) is an estimator based on an alternative form of the G-computation identifiability result defined in Equation (5) (Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983; Robins et al., 2000; Bodnar et al., 2004; Cole and Hernán, 2008). In this form, the statistical estimand is a function of the conditional probability of being exposed, given the adjustment covariates $\mathbb{P}(A=1|W)$, which is often called the *propensity score* (Rosenbaum and Rubin, 1983). IPW controls for confounding by up-weighting rare exposure-covariate subgroups, which have a small propensity score, and down-weighting more common subgroups, which have a larger propensity score. The IPW estimator is implemented with the following steps.

1. Regress the exposure on the covariate adjustment set to estimate the propensity score $\mathbb{P}(A=1|W)$.

2. Based on the estimates from Step 1, predict each individual's probability of receiving her observed exposure, given the adjustment covariates:
$$\hat{\mathbb{P}}(A_i|W_i) \text{ for all observations } i = 1, ..., N.$$
3. Obtain a point estimate by taking the empirical mean of the outcome weighted by the inverse of the conditional exposure probabilities:
$$\hat{\Psi}_{IPW}(\hat{\mathbb{P}}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1|W_i)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0|W_i)}\right)Y_i.$$
Thus, individuals who are exposed receive weight as one over the estimated propensity score $\hat{\mathbb{P}}(A_i = 1|W_i)$, while individuals who are not exposed receive weight as negative one over the estimated probability of not being exposed, given the covariates $\hat{\mathbb{P}}(A_i = 0|W_i)$.

The performance of the parametric G-computation depends on consistent estimation of the conditional expectation of the outcome, given the exposure and covariates $\mathbb{E}(Y|A,W)$, and the performance of IPW relies on consistent estimation of the propensity score $\mathbb{P}(A = 1|W)$. Traditionally, both estimators have relied on parametric regression models to estimate these quantities. If sufficient background knowledge is available to support using such a regression, it should have already been encoded in the causal model, yielding parametric structural equations in Section 2.2, and can be incorporated during estimation.

However, in most real-world studies with a large number of covariates and potentially complicated relationships, we usually do not have the knowledge support using such parametric regressions. More often, our statistical model $\mathscr{M}$ for the set of possible distributions of the observed data is non-parametric or semi-parametric (Section 2.4). Furthermore, we want to avoid introducing new and unsubstantiated assumptions during estimation. Reliance on poorly specified parametric regressions can result in biased point estimates and misleading inference (e.g., Benkeser et al. (2017); Luque-Fernandez et al. (2018)). At the same time, non-parametric methods, such as stratification, will break down due to sparsity. Here, recent advances in machine learning can help us estimate $\mathbb{E}(Y|A,W)$ and $\mathbb{P}(A = 1|W)$ without introducing new assumptions.

Data-adaptive estimation or machine learning techniques can be used to effectively estimate the *nuisance parameters*, which are the quantities needed to compute our statistical estimand: $\mathbb{E}(Y|A,W)$ and $\mathbb{P}(A = 1|W)$. We focus our discussion on *ensemble learning methods*, which "stack" or combine several prediction algorithms together and can be implemented as follows (Wolpert, 1992; Breiman, 1996).

First, we pre-specify a library of candidate algorithms, such as generalized linear models, splines, random forests, neural networks, or support vector machines. We also define a measure of performance through an appropriate loss function, such as the squared error or the negative log-likelihood. Next, we randomly split the observed data into V mutually exclusive and exhaustive sets of size N/V, called "folds". In each iteration, a single fold is chosen as a validation set and the remaining $V-1$ folds serve as the training set. We then fit each algorithm using only data from the training set and predict the outcomes for the units in the validation set. Each algorithm's performance is quantified by the average deviations, corresponding to the loss function, between the actual and predicted outcomes for the units in the validation set. Repeating the process V times, where each fold is used once as a validation set, amounts to performing *V-fold cross-validation*. We could then select the algorithm with the best performance, corresponding to the smallest average discrepancy with respect to the specified loss function.

This procedure, sometimes called Discrete Super Learner (van der Laan et al., 2007), effectively sets up a competition between the algorithms specified in the library, and selects the one with the best performance. It naturally follows then that Discrete Super Learner can only perform as well as the best-performing algorithm specified in its library. The full Super Learner algorithm improves upon its discrete version by taking a weighted combination of the algorithm-specific predictions to create a new prediction algorithm. We refer the reader to Polley et al. (2011) for further discussion of Super Learner and its properties and to Naimi and Balzer (2018) for worked examples and `R` code. The algorithm is available in the `SuperLearner` package in `R` (Polley et al., 2018).

The goal of Super Learner is to do the best possible job, according to the specified risk criterion, of predicting the outcome Y, given the exposure A and covariates W, or predicting the exposure A, given the covariates W. As a result, Super Learner-based estimators of the nuisance parameters $\mathbb{E}(Y|A,W)$ or $\mathbb{P}(A=1|W)$ have the wrong bias-variance tradeoff for our statistical estimand $\Psi(\mathbb{P})$, which is a single number as opposed to a whole prediction function. TMLE, discussed next, provides one way to integrate data-adaptive algorithms, such as Super Learner, and still obtain the best possible bias-variance tradeoff for the statistical estimand of interest. Indeed, a particular appeal of the Targeted Learning framework is the use of flexible estimation methods to respect the statistical model, which is often non-parametric, and to minimize the risk of bias due to regression model misspecification.

Targeted Maximum Likelihood Estimation (TMLE) provides a general approach to constructing double robust, semi-parametric, efficient, substitution estimators (van der Laan and Rubin, 2006; van der Laan and Rose, 2011; Petersen et al., 2014). Here, we provide a very brief overview and refer the reader to Schuler and Rose (2017) for a thorough introduction to the algorithm, which is available in the `tmle`, `ltmle`, and `drtmle` packages in `R` (Gruber and van der Laan, 2012; Lendle et al., 2017; Benkeser et al., 2017). To implement TMLE for the G-computation identifiability result $\Psi(\mathbb{P})$, given in Equation 4, we take the following steps.

First, we use Super Learner to compute an initial estimator of the conditional mean outcome, given the exposure and covariates $\hat{\mathbb{E}}^0(Y|A,W)$. Next, we "target" this initial estimator using information from the propensity score $\hat{\mathbb{P}}(A=1|W)$, also estimated with Super Learner. Informally, this targeting step can be thought of as a second chance to control confounding and serves to reduce statistical bias for the $\Psi(\mathbb{P})$. We denote the updated estimator of the conditional mean outcome as $\hat{\mathbb{E}}^1(Y|A,W)$ and use it to obtain targeted predictions of the outcome setting the exposures of interest, but keeping the covariates the same: $\hat{\mathbb{E}}^1(Y_i|A_i=1,W_i)$ and $\hat{\mathbb{E}}^1(Y_i|A_i=0,W_i)$ for all observations $i=1,\ldots,N$. Finally, we obtain a point estimate by taking the average difference in these targeted predictions.

$$\hat{\Psi}_{TMLE}(\hat{\mathbb{P}}) = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mathbb{E}}^1(Y_i|A_i=1,W_i) - \hat{\mathbb{E}}^1(Y_i|A_i=0,W_i)\right].$$

TMLE's updating step also serves to endow the algorithm with a number of theoretical properties, which often translate into superior performance in finite samples. First, under regularity and empirical process conditions detailed in van der Laan and Rose (2011), TMLE behave like maximum likelihood estimators in that the Central Limit Theorem can be invoked to study their limiting behavior; and so the normal distribution can be used for constructing confidence intervals and hypothesis testing, even if machine learning is used for estimation of the nuisance parameters $\mathbb{E}(Y|A,W)$ or $\mathbb{P}(A=1|W)$. Furthermore, the estimator is *double robust* in that it will be consistent if either $\mathbb{E}(Y|A,W)$ or $\mathbb{P}(A=1|W)$ is consistently estimated. Collaborative TMLE further improves upon this robustness result (van der Laan and Gruber, 2010; Gruber and van der Laan, 2015). Finally, if both nuisance parameters are estimated consistently and at fast enough rates, the estimator will be locally efficient and in large samples attain the minimal variance in a semi-parametric statistical model. We refer the reader to Kennedy (2017) for an introduction to semi-parametric efficiency theory.

Finally, we note that there is nothing inherent in the TMLE algorithm that demands the use of Super Learner. However, its implementation with machine learning algorithms avoids introducing new unsubstantiated assumptions during estimation and improve our chances for consistent results. Again, relying on misspecified parametric regressions can induce statistical bias and yield misleading statistical inference.

2.7. Interpretation of Results

The final step of the Roadmap is to interpret our results. We have seen that the causal inference framework clearly delineates the assumptions made from domain knowledge (Section 2.2) from the ones desired for identifiability (Section 2.5). In other words, this framework ensures that the assumptions needed to augment the statistical results with a causal interpretation are made explicit. In this regard, Petersen and van der Laan (2014) argue for a hierarchy of interpretations with "increasing strength of assumptions". First, we always have a statistical interpretation as an estimate of the difference in the expected outcome between exposed and unexposed units with the same covariate values, standardized over the covariate distribution in the population. We can also interpret $\hat{\Psi}(\hat{\mathbb{P}})$ as an estimate of the marginal difference in the expected outcome associated with the exposure, after controlling for measured confounding. To interpret our estimates causally, we need the identifiability assumptions (Section 2.5) to hold in the original causal model (Section 2.2). If either graphs in Figure 4 represented the true causal structure that generated our data and the positivity assumption held, then we could interpret $\hat{\Psi}(\hat{\mathbb{P}})$ as the average treatment effect or for a binary outcome the causal risk difference.

Now, recall that the counterfactual outcomes were derived through intervening on the causal model (Section 2.3). The selected intervention should match our underlying scientific question (Section 2.1) and does not have to correspond to a feasible or realistic intervention. If the identifiability assumptions (Section 2.5) held and the intervention could be conceivably implemented in the real world, then we could further interpret $\hat{\Psi}(\hat{\mathbb{P}})$ as an estimate of the intervention's impact if it had been implemented in the population of interest. Finally, if the identifiability assumptions were met and the intervention implemented perfectly in a study sample, whose characteristics exactly matched those of our population and who were fully measured, then we could interpret $\hat{\Psi}(\hat{\mathbb{P}})$ as replicating the results of the randomized trial of interest. We note this hierarchy represents a divergence from the Target Trial framework of Hernán and Robins (2016), who suggest causal inference with observational data can be thought of as "emulating" a randomized trial.

In our running example, the causal model shown in Figure 1 represents our knowledge of the data generating process; there are measured (W_1, W_2) as well as unmeasured U common causes of the exposure A and the outcome Y. Thus, the lack of identifiability prevents any interpretation as a causal effect or further along the hierarchy. Thus, we can interpret a point estimate of $\Psi(\mathbb{P})$ as the difference in the probability of HIV RNA viral suppression associated with pregnancy after controlling for the measured demographic and clinical confounders.

3. Conclusion

The objective of statistical analyses is to make inferences about the data generating process underlying a randomized trial or an observational study. In practice, statistical inference is concerned with purely data-driven tasks, such as prediction, estimation and hypothesis testing. In recent decades, the advent of causal inference has triggered a shift in focus, particularly within the data analysis community, toward a territory that has traditionally evaded statistical reach: the causal mechanism underlying a data generating process. Statistical inference relies on patterns present in the observed data, such as correlation, and therefore is unable, alone, to answer questions of causal nature (Pearl, 2010; Pearl et al., 2016). Nonetheless, questions about cause and effect are of prime importance in all fields including Data Science (Pearl, 2018; Hernán et al., 2018).

We have presented an overview of one framework for causal inference. We emphasized how the Causal Roadmap helps ensure consistency and transparency between the imperfect nature of real world data, and the complexity associated with questions of causal nature. Of course, this work serves only as a primer to causal inference in Data Science, and we have only presented the fundamental concepts and tools in the causal inference arsenal.

Indeed, this framework can be extended to richer and more complicated questions. For instance, our running example for average treatment effect only focused on a single exposure at a single time point. However, as demonstrated in Tran et al. (2016); Kreif et al. (2017), the Causal Roadmap can also handle multiple intervention nodes with time-dependent confounding. Other recent avenues of research in causal inference are discussed in the Appendix.

As a final note, a Data Scientist may debate the usefulness of applying the causal inference machinery to her own research. We hope to have clarified that if appropriately followed, the Causal Roadmap forces us to think carefully about the goal of our research, the context in which data were collected, and to explicitly define and justify any assumptions. It is our belief that conforming to the rigors of this causal inference framework will improve the quality and reproducibility of all scientific endeavors that rely on real data to understand how nature works.

Appendix

Here, we briefly highlight some extensions to more advanced settings. For each, we provide a broad definition and a few examples with citations to some relevant works.

1. **Marginal structural models** provide a summary of how the distribution of the counterfactual outcome changes as a function of the exposure and possibly pre-exposure covariates (Robins, 1999; Robins et al., 2000; Bodnar et al., 2004; Neugebauer and van der Laan,

2007; Robins and Hernán, 2009; Petersen and van der Laan, 2011; Zheng et al., 2016). Marginal structural models are another way to define our target causal parameter and especially useful when the exposure is continuous or has many levels.
Examples: Robins et al. (2000) specified a logistic regression model to summarize the dose-response relation for the cumulative effect of zidovudine (AZT) treatment on the counterfactual risk of having undetectable HIV RNA levels among HIV-positive patients. For a time-to-event outcome, Cole et al. (2012) used a Cox proportional hazard model to summarize the association between treatment initiation and the counterfactual hazard of incident AIDS or death among persons living with HIV.

2. **Longitudinal exposures**, corresponding to interventions on multiple treatment nodes, allow us to assess the cumulative effect of an exposure or exposures over time (Robins et al., 2000; Bang and Robins, 2005; Robins and Hernán, 2009; Petersen and van der Laan, 2011; van der Laan and Gruber, 2012; Westreich et al., 2012; Petersen et al., 2014). Examining the effects of longitudinal exposures is complicated by time-dependent confounding, when a covariate is affected by a prior treatment and confounds a future treatment. In these settings, causal frameworks have been especially useful for identifying the appropriate adjustment sets and thereby statistical analysis.
Examples: Schnitzer et al. (2014) sought to assess the effect of breastfeeding duration on gastrointestinal infections among new borns, while Decker et al. (2014) investigated the effects of sustained physical activity and diet interventions on adolescent obesity.

3. **Effect mediation** refers to a general class of causal questions seeking to distinguish an exposure's direct effect on the outcome from its indirect effect through an intermediate variable (Robins and Rotnitzky, 1992; Pearl, 2001; Petersen et al., 2006; van der Laan and Petersen, 2008; VanderWeele, 2009; Imai et al., 2010; Zheng and van der Laan, 2012; Tran et al., 2016). There are several types of direct and indirect effects. For example, the controlled direct effect refers to the contrast between the expected counterfactual outcomes under two levels of the exposure, but when the mediator is fixed at a constant level. The natural direct effect, also called the pure direct effect, refers to the contrast between the expected counterfactual outcomes under two levels of the exposure, but when the mediator remains at its counterfactual level under the reference value of the exposure. Indirect effects can be defined analogously.
Examples: Naimi et al. (2016) examined the disparity in infant mortality due to race that would remain if all mothers breastfeed prior to hospital discharge. More recently, Rudolph et al. (2018) investigated how the impact of neighborhood disadvantage on adolescent substance use was mediated by school and peer environment.

4. **Dynamic treatment regimes** are personalized rules for assigning the exposure or treatment as a function of an individual's covariate history (Murphy, 2003; Hernán et al., 2006; van der Laan and Petersen, 2007; Kitahata et al., 2009; Hernán and Robins, 2009; Cain et al., 2010; Kreif et al., 2017). They are also called "adaptive treatment strategies" and "individualized treatment rules". Static interventions, which assign a single level of the exposure to all individuals regardless of their covariate values, can be considered a special case of dynamic interventions.
Examples: Cain et al. (2010) and Young et al. (2011) both considered CD4-based thresh-

olds for initiating antiretroviral therapy and their impact on mortality among persons living with HIV. Recently, Kreif et al. (2017) compared static and dynamic regimes to understand the optimal timing and level of nutritional support for children in a pediatric intensive care unit.

5. **Stochastic interventions** aim to change or shift the distribution of the exposure (Korb et al., 2004; Taubman et al., 2009; Cain et al., 2010; Díaz and van der Laan, 2012, 2013a; Rudolph et al., 2017). Stochastic interventions are especially useful when the exposure of interest can not be directly manipulated and can help alleviate violations to the positivity assumption. Deterministic interventions, which assign a given level of the exposure with probability one, can be considered a special case of stochastic interventions.
   Examples: Díaz and van der Laan (2012) asked what is the impact of a policy encouraging more exercise, according to health and socioeconomic factors, on mortality in a population of older adults? Danaei et al. (2013) examined the impact of various lifestyle interventions, such as eating at least 2 servings of whole grain per day, on the risk of type 2 diabetes in women.

6. **Clustered data** occur when there is dependence or correlation between individuals within some grouping, such as a clinic, school, neighborhood, or community. Such correlation can arise from shared cluster-level factors, including the exposure, and from social or biological interactions between individuals with a cluster (Halloran and Struchiner, 1991, 1995; Oakes, 2004; Tchetgen Tchetgen and VanderWeele, 2012; van der Laan, 2014; Schnitzer et al., 2014; Prague et al., 2016; Balzer et al., 2018; Morozova et al., 2018; Buchanan et al., 2018). This dependence must be accounted when specifying the causal model and often demands relaxing the stable unit treatment value assumption, which prohibits one unit's exposure from impacting another's outcome (Rubin, 1978).
   Examples: Balzer et al. (2018) examined the impact of household socioeconomic status, a cluster-level variable, on the risk of failing to test for HIV. Likewise, Buchanan et al. (2018) investigated both the individual and disseminated effects of a network-randomized intervention among people who inject drugs.

7. **Missing data, censoring, and losses to follow up** can all be treated as additional intervention nodes in a given causal framework (Robins et al., 2000, 1994; Scharfstein et al., 1999; Daniel et al., 2012; Mohan et al., 2013; Balzer et al., 2017). Thereby, we can treat missing data as a causal inference problem - as opposed to causal inference as a missing data.
   Examples: When estimating the effect of iron supplementation during pregnancy on anemia at delivery, Bodnar et al. (2004) used inverse probability of censoring weights to adjust for the measured ways in which the women who were censored could differ from those who were not. Likewise, Petersen et al. (2017) estimated the probability of HIV RNA viral suppression over time among a closed cohort of HIV-infected adults, under a hypothetical intervention to prevent censoring and ensure complete viral load measurement.

8. **Transportability**, a subset of generalizability, aims to apply the effect for a given sample to a different population or setting (Cole and Stuart, 2010; Stuart et al., 2011; Hernán and VanderWeele, 2011; Petersen, 2011; Bareinboim and Pearl, 2013; Pearl, 2015; Lesko et al., 2017; Balzer, 2017).

Examples: Rudolph and van der Laan (2017) examined whether the reduction in school dropout observed in the Moving to Opportunity trial was consistent between Boston and Los Angeles. Recently, Hong et al. (2018) investigated whether the reductions in cardiovascular risk from rosuvastatin as observed in the JUPITER trial would also have been observed in the UK population who were trial eligible.

References

Ahern, J. (2018). Start with the "C-word," follow the roadmap for causal inference. *American Journal of Public Health*, 108(5):621.

Balzer, L. (2017). “All generalizations are dangerous, even this one.” - Alexandre Dumas [Commentary]. *Epidemiology*, 28(4):562–566.

Balzer, L., Petersen, M., and van der Laan, M. (2016). Tutorial for causal inference. In Buhlmann, P., Drineas, P., Kane, M., and van der Laan, M., editors, *Handbook of Big Data*. Chapman & Hall/CRC.

Balzer, L., Schwab, J., van der Laan, M., and Petersen, M. (2017). Evaluation of progress towards the UNAIDS 90-90-90 HIV care cascade: A description of statistical methods used in an interim analysis of the intervention communities in the SEARCH study. Technical Report 357, University of California at Berkeley. http://biostats.bepress.com/ucbbiostat/paper357/.

Balzer, L., Zheng, W., van der Laan, M., Petersen, M., and the SEARCH Collaboration (2018). A new approach to hierarchical data analysis: Targeted maximum likelihood estimation for the causal effect of a cluster-level exposure. *Stat Meth Med Res*, OnlineFirst.

Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972.

Bareinboim, E. and Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134.

Benkeser, D., Carone, M., van der Laan, M., and Gilbert, P. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880.

Bodnar, L., Davidian, M., Siega-Riz, A., and Tsiatis, A. (2004). Marginal Structural Models for Analyzing Causal Effects of Time-dependent Treatments: An Application in Perinatal Epidemiology. *American Journal of Epidemiology*, 159(10):926–934.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24:49–64.

Buchanan, A., Vermund, S., Friedman, S., and Spiegelman, D. (2018). Assessing individual and disseminated effects in network-randomized studies. *Am J Epidemiol*, 187(11):2449–2459.

Cain, L., Robins, J., Lanoy, E., Logan, R., Costagliola, D., and Hernán, M. (2010). When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics*, 6(2):Article 18.

Cole, S. and Hernán, M. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

Cole, S., Hudgens, M., Tien, P., Anastos, K., Kingsley, L., Chmiel, J., and Jacobson, L. (2012). Marginal structural models for case-cohort study designs to estimate the association of antiretroviral therapy initiation with incident AIDS or death. *Am J Epidemiol*, 175(5):381–390.

Cole, S. and Stuart, E. (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 Trial. *American Journal of Epidemiology*, 172(1):107–115.

Danaei, G., Pan, A., Hu, F., and Hernán, M. (2013). Hypothetical midlife interventions in women and risk of type 2 diabetes. *Epidemiol*, 24(1):122–128.

Daniel, R., Kenward, M., Cousens, S., and De Stavola, B. (2012). Using causal diagrams to guide analysis in missing data problems. *Stat Meth Med Res*, 21(3):243–256.

Dawid, A. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

Decker, A., Hubbard, A., Crespi, C., Seto, E., and Wang, M. (2014). Semiparametric estimation of the impacts of longitudinal interventions on adolescent obesity using targeted maximum-likelihood: Accessible estimation with the ltmle package. *Journal of Causal Inference*, 2(1):95–108.

Descartes, R. (1637). *Discours de la Méthode Pour bien conduire sa raison, et chercher la vérité dans les sciences*. Leiden, Netherlands.

Díaz, I. and van der Laan, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549.

Díaz, I. and van der Laan, M. (2013a). Assessing the causal effect of policies: An example using stochastic interventions. *Int J Biostat*, 9(2):161–174.

Díaz, I. and van der Laan, M. (2013b). Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *Int J Biostat*, 9:149–160.

Duncan, O. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.

Goldberger, A. (1972). Structural equation models in the social sciences. *Econometrica: Journal of the Econometric Society*, 40:979–1001.

Gruber, S. and van der Laan, M. (2012). tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(13):1–35.

Gruber, S. and van der Laan, M. (2015). Consistent causal effect estimation under dual misspecification and implications for confounder selection procedures. *Stat Methods Med Res*, 24(6):1003–1008. PMID: 22368176.

Halloran, M. and Struchiner, C. (1991). Study designs for dependent happenings. *Epidemiology*, 2:331–338.

Halloran, M. and Struchiner, C. (1995). Causal inference in infectious diseases. *Epidemiology*, 6(2):142–151.

Heckman, J. and Vytlacil, E. (2007). Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. *Handbook of Econometrics*, pages 4779–4874.

Hernán, M. (2005). Invited commentary: hypothetical interventions to define causal effects–afterthought or prerequisite? *Am J Epidemiol*, 162(7):618–620.

Hernán, M. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, 108(5):616–619.

Hernán, M., Alonso, A., Logan, R., Grodstein, F., Michels, K., Willett, W., Manson, J., and Robins, J. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, 19:766–779.

Hernán, M., Hernández-Díaz, S., and Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Hernán, M., Hsu, J., and Healy, B. (2018). Data science is science's second chance to get causal inference right: A classification of data science tasks. Technical report, arXiv. https://arxiv.org/abs/1804.10846.

Hernán, M., Lanoy, E., Costagliola, D., and Robins, J. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology*, 98(3):237–242.

Hernán, M. and Robins, J. (2009). Comment on: Early versus deferred antiretroviral therapy for HIV on survival. *New England Journal of Medicine*, 361(8):823–824.

Hernán, M. and Robins, J. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764.

Hernán, M. and VanderWeele, T. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, 22:368–377.

Hernández-Díaz, S., Schisterman, E., and Hernán, M. (206). The birth weight "paradox" uncovered? *Am J Epidemiol*, 164(11):1115–1120.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hong, J., Jonsson Funk, M., LoCasale, R., Dempster, S., Cole, S., Webster-Clark, M., Edwards, J., and Sturmer, T. (2018). Generalizing randomized clinical trial results: Implementation and challenges related to missing data in the target population. *Am J Epidemiol*, 184(4):817–827z.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25:51–71.

Joint United Nations Programme on HIV/AIDS (UNAIDS) (2014). The gap report. Geneva, Switzerland.

Kennedy, E. (2017). Semiparametric theory. Technical report, arXiv. https://arxiv.org/abs/1709.06418v1.

Kitahata, M., Gange, S., Abraham, A., Merriman, B., Saag, M., Justice, A., et al. (2009). Effect of early versus deferred antiretroviral therapy for HIV on survival. *New England Journal of Medicine*, 360(18):1815–1826.

Korb, K., Hope, L., Nicholson, A., and Axnick, K. (2004). Varieties of causal intervention. In Zhang, C., Guesgen, H., and Yeap, W., editors, *PRICAI 2004: Trends in Artificial Intelligence, volume 3157 of Lecture Notes in Computer*

Science, pages 322–331. Springer, Heidelberg, Germany.

Kreif, N., Tran, L., Grieve, R., De Stavola, B., Tasker, R., and Petersen, M. (2017). Estimating the comparative effectiveness of feeding interventions in the pediatric intensive careunit: A demonstration of longitudinal targeted maximum likelihood estimation. *American Journal of Epidemiology*, 186(12):1370–1379.

Lendle, S., Schwab, J., Petersen, M., and van der Laan, M. (2017). ltmle: An R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, 81(1):1–21.

Lesko, C., Buchanan, A., Westreich, D., Edwards, J., Hudgens, M., and Cole, S. (2017). Generalizing study results: a potential outcomes perspective. *Epidemiology*, 28(4):553–561.

Little, R. and Rubin, D. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Revue of Public Health*, 21:121–145.

Luque-Fernandez, M., Belot, A., Valeri, L., Cerulli, G., Maringe, C., and Rachet, B. (2018). Data-adaptive estimation for double-robust methods in population-based cancer epidemiology: Risk differences for lung cancer mortality by emergency presentation. *American Journal of Epidemiology*, 187(4):871–878.

Marcus, G. and Davis, E. (2014). Eight (no, nine!) problems with big data. *The New York Times*.

Messer, L., Oakes, J., and Mason, S. (2010). Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *American Journal of Epidemiology*, 171:664–673.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc.

Morozova, O., Cohen, T., and Crawford, F. (2018). Risk ratios for contagious outcomes. *J. R. Soc. Interface*, 15(20170696).

Murphy, S. (2003). Optimal dynamic treatment regimes. *J R Stat Soc Ser B*, 65(2):331–355.

Naimi, A. and Balzer, L. (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, pages 459–464.

Naimi, A., Schnitzer, M., Moodie, E., and Bodnar, L. (2016). Mediation analysis for health disparities research. *Am J Epidemiol2016*, 184(4):315–324.

Neugebauer, R. and van der Laan, M. J. (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes (In Polish). English translation by D.M. Dabrowska and T.P. Speed (1990). *Statistical Science*, 5:465–480.

Oakes, J. (2004). The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology (with discussion). *Soc Sci Med*, 58(10):1929–1952. PMID: 15020009.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:669–710.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York. Second ed., 2009.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420, San Francisco. Morgan Kaufmann.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):Article 7.

Pearl, J. (2015). Generalizing experimental findings. *Journal of Causal Inference*, 3(2):259–266.

Pearl, J. (2018). The seven tools of causal inference with reflections on machine learning. Technical Report R-481, UCLA.

Pearl, J., Glymour, M., and Jewell, N. (2016). *Causal inference in statistics: a primer*. John Wiley and Sons Ltd, Chichester, West Sussex, UK.

Petersen, M. (2011). Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. *Epidemiology*, 22:378–381.

Petersen, M. and Balzer, L. (2014). Introduction to causal inference. UC Berkeley. www.ucbbiostat.com/labs.

Petersen, M., Balzer, L., Kwarsiima, D., Sang, N., et al. (2017). Association of implementation of a universal testing and treatment intervention with HIV diagnosis, receipt of antiretroviral therapy, and viral suppression among adults in East Africa. *JAMA*, 317(21):2196–2206.

Petersen, M., LeDell, E., Schwab, J., Sarovar, V., Gross, R., Reynolds, N., et al. (2015). Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. *J Acquir Immune Defic Syndr*, 69(1):109–118.

Petersen, M., Porter, K., Gruber, S., Wang, Y., and van der Laan, M. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54.

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., and van der Laan, M. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2(2).

Petersen, M., Sinisi, S., and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology*, 17(3):276–284.

Petersen, M. and van der Laan, M. (2011). Case Study: Longitudinal HIV Cohort Data. In van der Laan, M. and Rose, S., editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London.

Petersen, M. and van der Laan, M. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, 25(3):418–426.

Polley, E., LeDell, E., Kennedy, C., and van der Laan, M. (2018). *SuperLearner: Super Learner Prediction*. R package version 2.0-24.

Polley, E., Rose, S., and van der Laan, M. (2011). Super Learner. In van der Laan, M. and Rose, S., editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London.

Prague, M., Wang, R., Stephens, A., E. Tchetgen Tchetgen, and De Gruttola, V. (2016). Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. *Biometrics*, 72(4):1066–1077.

Richardson, T. and Robins, J. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Working paper number 128, Center for Statistics and the Social Sciences University of Washington.

Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512.

Robins, J. (1999). Association, Causation, and Marginal Structural Models. *Synthese*, 121(1-2):151–179.

Robins, J. and Hernán, M. (2009). Estimation of the causal effects of time-varying exposures. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, chapter 23. Chapman & Hall/CRC, Boca Raton, FL.

Robins, J., Hernán, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In Jewell, N., Dietz, K., and Farewell, V., editors, *AIDS Epidemiology - Methodological Issues*, Boston. Birkhäuser.

Robins, J., Rotnitzky, A., and Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Halloran, M. and Berry, D., editors, *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Springer, New York.

Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Rose, S. (2012). Big data and the future. *Significance*, 9(4):47–48.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies. *Biometrika*, 70:41–55.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann Stat*, 6:34–58.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rudolph, K., Goin, D., Paksarian, D., Crowder, R., Merikangas, K., and Stuart, E. (2018). Causal mediation analysis with observational data: Considerations and illustration examining mechanisms linking neighborhood poverty to adolescent substance use. *Am J Epidemiol*, Epub Ahead of Print.

Rudolph, K., Sofrygin, O., Zheng, W., and van der Laan, M. (2017). Robust and flexible estimation of stochastic mediation effects: a proposed method and example in a randomized trial setting. *Epidemiologic Methods*, 7.

Rudolph, K. and van der Laan, M. (2017). Robust estimation of encouragement-design intervention effects transported across sites. *J R Stat Soc Ser B*, 79(5):1509–1525.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adjusting for Nonignorable Drop-Out Using Semiparametric

Nonresponse Models (with Rejoiner). *Journal of the American Statistical Association*, 94(448):1096–1120 (1135–1146).

Schnitzer, M., van der Laan, M., Moodie, E., and Platt, R. (2014). Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data. *Annals of Applied Statistics*, 8(2):703–725.

Schuler, M. and Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search. Number 81 in Lecture Notes in Statistics*. Springer-Verlag, New York/Berlin.

Stuart, E., Cole, S., Bradshaw, C., and Leaf, P. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A*, 174(Part 2):369–386.

Taubman, S., Robins, J., Mittleman, M., and Hernán, M. (2009). Intervening on risk factors for coronary heart disease: an application of the parametric G-formula. *International Journal of Epidemiology*, 38(6):1599–1611.

Tchetgen Tchetgen, E. and VanderWeele, T. (2012). On causal inference in the presence of interference. *Stat Meth Med Res*, 21(1):55–75.

Tran, L., Yiannoutsos, C., Musick, B., Wools-Kaloustian, K., Siika, A., Kimaiyo, S., van der Laan, M., and Petersen, M. (2016). Evaluating the impact of a HIV low-risk express care task-shifting program: A case study of the targeted learning roadmap. *Epidemiologic Methods*, 5(1):69–91.

van der Laan, M. (2014). Causal inference for a population of causally connected units. *Journal of Causal Inference*, 0(0):1–62.

van der Laan, M. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

van der Laan, M. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1).

van der Laan, M., Haight, T., and Tager, I. (2005). van der Laan et al. respond to "hypothetical interventions to define causal effects”. *Am J Epidemiol*, 162(7):621–622.

van der Laan, M. and Petersen, M. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1):Article 3.

van der Laan, M. and Petersen, M. (2008). Direct effect models. *The International Journal of Biostatistics*, 4(1):Article 23.

van der Laan, M., Polley, E., and Hubbard, A. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):25.

van der Laan, M. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London.

van der Laan, M. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11.

VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26.

VanderWeele, T. and Arah, O. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 22:42–52.

Westreich, D., Cole, S., Young, J., Palella, F., Tien, P., Kingsley, L., Gange, S., and Hernán, M. (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident aids or death. *Statistics in Medicine*, 31(18):2000–2009.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.

Young, J., Cain, L., Robins, J., O’Reilly, E., and Hernán, M. (2011). Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula. *Stat Biosci*, 3:119–143.

Zhang, Y., Young, J., Thamer, M., and Hernán, M. (2018). Comparing the effectiveness of dynamic treatment strategies using electronic health records: An application of the parametric g-formula to anemia management strategies. *Health Serv Res*, 53(3):1900–1918.

Zheng, W., Petersen, M., and van der Laan, M. (2016). Doubly robust and efficient estimation of marginal structural models for the hazard function. *Int J Biostat*, 12(1):233–252.

Zheng, W. and van der Laan, M. (2012). Targeted maximum likelihood estimation for natural direct effects. *The International Journal of Biostatistics*, 8(1):1–40.