The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 23

Direct Effect Models

Mark J. van der Laan, *University of California, Berkeley*
Maya L. Petersen, *University of California, Berkeley*

Recommended Citation:
van der Laan, Mark J. and Petersen, Maya L. (2008) "Direct Effect Models," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 23.
DOI: 10.2202/1557-4679.1064

Direct Effect Models

Mark J. van der Laan and Maya L. Petersen

Abstract

The causal effect of a treatment on an outcome is generally mediated by several intermediate variables. Estimation of the component of the causal effect of a treatment that is not mediated by an intermediate variable (the direct effect of the treatment) is often relevant to mechanistic understanding and to the design of clinical and public health interventions. Robins, Greenland and Pearl develop counterfactual definitions for two types of direct effects, natural and controlled, and discuss assumptions, beyond those of sequential randomization, required for the identifiability of natural direct effects. Building on their earlier work and that of others, this article provides an alternative counterfactual definition of a natural direct effect, the identifiability of which is based only on the assumption of sequential randomization. In addition, a novel approach to direct effect estimation is presented, based on assuming a model directly on the natural direct effect, possibly conditional on a subset of the baseline covariates. Inverse probability of censoring weighted estimators, double robust inverse probability of censoring weighted estimators, likelihood-based estimators, and targeted maximum likelihood-based estimators are proposed for the unknown parameters of this novel causal model.

KEYWORDS: causal inference, counterfactual, double robust estimation, G-computation, inverse probability of treatment/censoring weighted estimation, targeted maximum likelihood

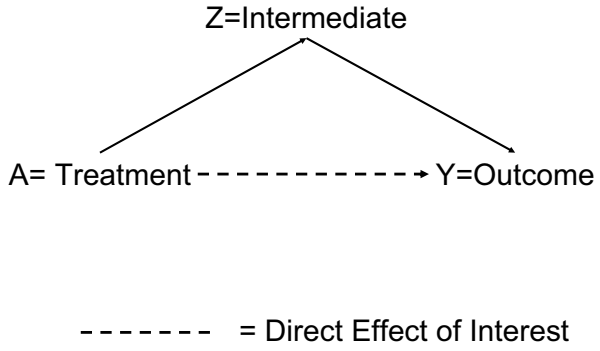
1 Introduction.

Epidemiological and clinical research often aims to estimate the causal effect of a treatment on an outcome. Definition of the counterfactual, or potential outcome, framework allowed causal inference to be framed as a missing data problem (Rubin (1978)), leading to major methodological advances. Under the counterfactual framework, the identifiability of causal parameters relies on coarsening at random (also known as sequential randomization or the assumption of no unmeasured confounders) (Rubin (1976); Neyman (1990); Heitjan and Rubin (1991); Jacobsen and Keiding (1995); Gill et al. (1997)). van der Laan and Robins (2002a) provide an overview of various causal models developed in this framework and of the corresponding literature.

It is often of considerable interest to identify the pathways by which a treatment is acting and to quantify the component of a treatment's effect is and is not mediated by a given intermediate variable (the indirect and direct effects of the treatment, respectively) is of considerable interest. Estimation of the direct and indirect effects of a treatment can inform mechanistic understanding of the treatment's action and the design of clinical and public health interventions. For example, treatment of Human Immunodeficiency Virus (HIV) infection using antiretroviral therapy suppresses viral replication, reflected in a reduced plasma HIV RNA level (viral load). As a result of reduced viral replication, a patient's CD4+ T lymphocyte count increases, restoring immunologic function. Antiretroviral therapy may also increase CD4+ T lymphocyte count in ways not mediated by changes in viral load (Deeks et al. (2000)). Estimation of the direct causal effect of antiretroviral therapy on CD4 count (not mediated by changes in viral load) has implications for understanding both the mechanics of antiretroviral action and the appropriate clinical response to viral resistance, which can reduce or eliminate the effect of treatment on viral load.

Robins and Greenland (1992) and Pearl (2000) use the counterfactual framework to develop definitions for direct effects and to address the identification and estimation of these effects. In defining direct effects, these authors assume, for a randomly sampled subject, the existence of counterfactual outcomes Y_{az} and counterfactual intermediate variables Z_a under 'set' values of the treatment $A = a$ and the intermediate variable $Z = z$. The observed data is viewed as a missing data structure on these counterfactuals, and two definitions of an individual direct effect are developed. Under the first definition an individual direct effect is defined as $Y_{aZ_0} - Y_{0Z_0}$, equivalent to the counterfactual effect of treatment $A = a$ on outcome Y when the intermediate variable is set at the counterfactual value Z_0 that would have been observed had the

Figure 1: Directed Acyclic Graph showing direct effect



individual received the reference level of treatment ($A = 0$). Under the second definition an individual direct effect is defined as $Y_{az} - Y_{0z}$, equivalent to the counterfactual effect of treatment $A = a$ on outcome Y when the intermediate variable is held constant at a (user-specified) set level. Here we follow the lead of Pearl (2000) and refer to the first type of direct effect as *natural* and the second as *controlled* (noting that Robins and Greenland (1992) refer to natural direct effects as "pure direct effects"). Population direct effects are defined as the mean of these individual counterfactual direct effects. Figure (1) shows a directed acyclic graph representing a direct effect.

Robins and Greenland (1992) introduce an additional assumption, beyond that of sequential randomization, needed for natural direct effects to be identifiable from the observed data; Robins and Greenland's No Interaction Assumption states that the individual controlled direct effect does not depend on the level at which the intermediate variable is fixed. Under this assumption, natural and controlled direct effects are equivalent; thus the authors discuss estimation of a single type of direct effect. Pearl (2000) provides an alternative identifying assumption for natural direct effects; he assumes that an individual's outcome under a fixed level of the treatment and intermediate variable does not depend on the counterfactual level of the intermediate under no treatment. Finally, van der Laan and Petersen (2004), introduce a third identifying assumption which states that, within strata of baseline covariates, the individual controlled direct causal effect is independent (in the mean sense) of the no-treatment counterfactual intermediate variable. As discussed by Robins (2003), van der Laan and Petersen (2004), and Petersen et al. (2006), all three

DOI: 10.2202/1557-4679.1064

of these identifying assumptions may prove unrealistic for some applications in epidemiology and clinical medicine.

In this article we first note that, in the absence of any identifying assumptions beyond that of sequential randomization, the parameter targeted by the identifiability result for the natural direct effect $E(Y_{aZ_0} - Y_{0Z_0})$ is still both interesting and interpretable. Specifically, it equals the population mean of a subject-specific average of z-specific controlled direct effects, $Y_{az} - Y_{0z}$, where the subject-specific average is obtained with respect to a conditional distribution of Z_0 given the subject's baseline covariates. The article focuses on modelling and estimation of this natural direct effect parameter, which happens to agree with the conventional natural direct effect parameter $E(Y_{aZ_0} - Y_{0Z_0})$ under any of the identifying assumptions of Robins and Greenland (1992), van der Laan and Petersen (2004) or Pearl (2000), but does not rely on any of these additional assumptions. In addition, the natural direct effect parameter is generalized by allowing the user to specify the choice of conditional distribution used to obtain an average of the z-specific controlled individual direct effects.

The article further introduces a novel approach to natural direct effect estimation. The approach is based on assuming a model for the natural direct effect parameter that makes as few assumptions as possible beyond correct specification under the null hypothesis. By the curse of dimensionality, some further modelling assumptions *are* typically necessary to obtain estimators with good practical performance; however, dependence on these assumptions can be minimized by applying the locally efficient estimating function-based methodology presented by van der Laan and Robins (2002a). Using this methodology, we develop inverse probability of censoring weighted (IPCW) estimators and double robust inverse probability of censoring weighted (DR-IPCW) estimators, where the latter have only second-order dependence on nuisance parameters.

The general approach to direct effect estimation presented in this chapter differs from previous approaches to the estimation of direct effects, such as those based on the identifiability result for natural direct effects (discussed in van der Laan and Petersen (2004); Petersen et al. (2006)), or based on inverse probability weighting (Robins, November 1999). These previous approaches to estimation involve specifying a model for the dependence of the counterfactual outcome on the treatment and intermediate variables; depending on the approach used, specifying the dependence of the counterfactual outcome on the treatment, the intermediate and all confounding covariates may be required. In contrast, the approach presented here is based on assuming a model only for the direct effect parameter being estimated (a less parametric approach). We

note that the IPCW and DR-IPCW estimators of the direct effect introduced here are analogous to the marginal structural model estimators introduced by Robins (e.g., Robins (2000); Robins and Rotnitzky (2001)) in that they are developed using the same estimating function methodology. However, the estimators presented here are developed under a distinct causal model and estimate a distinct causal parameter (the natural direct effect).

The article is organized as follows. In Section 2 we present our proposed model for direct effects, which is based on the statistical counterfactual framework and assumes sequential randomization to ensure identifiability. In Section 3 we present and compare three methods for estimation, based, respectively, on inverse probability weighting, double-robust inverse probability weighting, and likelihood-based estimation of the identifiability result. Implementation is discussed and the three methods are compared. Section 4 introduces a 4th method for estimation of the direct effect parameter, based on targeted maximum likelihood estimation. Section 5 discusses statistical inference for the direct effect parameter, and Section 6 closes with a discussion. In this article we present models and corresponding estimators of the direct effect of a treatment that is assigned at a single point in time, followed by an intermediate variable measured at a single point in time .Formal generalization of the statistical models and estimators presented here to longitudinal data structures, in which both treatment regimens and intermediate processes can be time-dependent, is provided in an online technical report (van der Laan and Petersen (2005)).

2 Direct effect models.

Consider a longitudinal study in which one collects on n randomly sampled subjects the chronological data structure $O = (W, A, Z, Y)$, where W is a vector of baseline covariates measured before the initiation of treatment A, and Z is an intermediate variable of interest on the causal pathway from treatment to the final outcome Y (Figure 1). Let P_0 denote the sampling distribution of O.

2.1 Definition of direct effects.

In this subsection, we present the statistical framework and definitions of direct effects as presented by Robins and Greenland (1992) and Robins (2003), and followed in van der Laan and Petersen (2004) and Petersen et al. (2006). This statistical framework represents the observed data on a randomly sam-

DOI: 10.2202/1557-4679.1064

pled individual as a missing data structure, where the full data structure is a collection of counterfactual data structures corresponding with 'set' values of the treatment and intermediate variables. Specifically, the full data structure consists of the value of the intermediate variable resulting from each possible treatment, and the value of the outcome resulting from each possible combination of treatment and intermediate. The observed data structure is a subset of this full data structure, consisting of a single treatment and the corresponding intermediate variable and outcome.

Formally, we assume the existence of a random variable $X \equiv [(Z_a : a \in \mathcal{A}), \{Y_{az} : (a,z) \in \mathcal{B}\}]$ of treatment-specific counterfactuals Z_a and Y_{az} (for the randomly sampled subject). Here $\mathcal{A}$ and $\mathcal{B}$ denote the support of A and (A,Z), respectively. We further assume that

$$O = (W, A, Z = Z_A, Y = Y_{AZ}) \tag{1}$$

is a missing data structure on the full data structure X. That is, X is the full data structure of interest, A is the treatment (which acts as a missingness variable on the full data), and O is a specified function of X and A. The density of O can be factorized as

$$P(W, A, Z, Y) = P(W)P(A \mid W)P(Z \mid A, W)P(Y \mid W, A, Z).$$

Because O is a function of A and X, its distribution can be parameterized by the probability distribution $g(\cdot \mid X)$ of A, given X (called the treatment mechanism), and the distribution F_X of X. Thus $P_0 = P_{F_{X0},g_0}$ for some F_{X0} and g_0.

One defines the natural direct effect of changing treatment from 0 (representing a reference treatment or no treatment) to a within strata of the sampling population defined by a baseline covariate $V \subset W$ as

$$E(Y_{aZ_0} - Y_{0Z_0} \mid V).$$

Note that V is chosen to define the strata of interest in which the investigator wishes to estimate the natural direct effect.

In order to identify the controlled direct effect $E(Y_{az} - Y_{0z} \mid V)$, or the direct effect at a set level z of the intermediate, together with previous authors (for example, Robins and Greenland (1992); Poole and Kaufman (2000); Cole and Hernan (2002)) we make the following randomization assumption:

$$(A, Z) \perp (Y_{az} : a, z) \mid W. \tag{2}$$

or, equivalently

$$A \perp (Y_{az} : a, z) | W, \tag{3}$$

$$Z \perp (Y_{az} : a, z)|A, W. \tag{4}$$

In order to identify a conditional distribution of Z_a, given W, we assume

$$A \perp (Z_a : a \in \mathcal{A}) \mid W. \tag{5}$$

These randomization assumptions will be met if A and Z are assigned randomly. However, this will rarely be possible for Z, and in many cases will not be possible for A. Alternatively, one must assume (based on background knowledge) that sufficient covariates are measured to control for confounding of the effect of treatment on outcome, treatment on intermediate, and intermediate on outcome.

Because of these randomization assumptions (5) and (2), we have the following relation between observed data probabilities and counterfactual probabilities:

$$\begin{aligned} P(A = a, Z = z \mid W) &= P\{A = a, Z = z \mid (Y_{az} : a, z), W\} \\ P(Z = z \mid A = a, W) &= P(Z_a = z \mid W) \\ P(Y = y \mid W, A = a, Z = z) &= P(Y_{az} = y \mid W). \end{aligned}$$

Consider also the conditional independence assumption (in the mean sense) of van der Laan and Petersen (2004):

$$E(Y_{az} - Y_{0z} \mid Z_0 = z, W) = E(Y_{az} - Y_{0z} \mid W) \text{ for all } (a, z) \in \mathcal{B}. \tag{6}$$

In the above model for the observed data distribution defined by (1), (2), (5), and (6), van der Laan and Petersen (2004) show that $E(Y_{aZ_0} - Y_{0Z_0} \mid V)$ equals

$$E_{W|V} \int_z \{E(Y \mid A = a, Z = z, W) - E(Y \mid A = 0, Z = z, W)\} P(Z = z \mid A = 0, W), \tag{7}$$

and thus that $E(Y_{aZ_0} - Y_{0Z_0} \mid V)$ is a (non-parametric) identifiable parameter. As mentioned in the introduction and discussed in van der Laan and Petersen (2004), both Robins and Greenland (1992) and Pearl (2000) derive the identical identifiability result under slightly different assumptions.

2.2 A generalized class of direct effect parameters.

We argue that, even without the identifiability assumption (6), (7) is still a potentially interesting direct effect parameter since, by the randomization assumptions only, (7) equals

$$DE(a, V) \equiv E\{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid V\}. \tag{8}$$

DOI: 10.2202/1557-4679.1064

That is, it equals the conditional expectation, given V, of a subject-specific average, $\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W)$, of the z-specific individual controlled direct effects $Y_{az} - Y_{0z}$, averaged with respect to the conditional distribution of Z_0 given W. Therefore, if one is not comfortable with the identifiability assumption (6), then one can view the latter direct effect parameter $DE(a, V)$ as the parameter of interest, and our proposed estimators are estimators of $DE(a, V)$.

As a side note, we point out that Robins (2003) provides yet another alternative definition of a natural direct effect that does not rely on an additional identifying assumption. Specifically, he observes that $E_W\{\sum_z E(Y_{az} \mid W)\} P(Z_0 = z \mid W)$ always has the interpretation of the counterfactual mean of Y when A is set to a and Z is randomly assigned according to the distribution of Z in the absence of treatment given W; the population natural direct effect can be viewed as the difference in the mean of this counterfactual outcome and the mean of the counterfactual outcome in the absence of treatment.

This direct effect definition (8) can be further generalized to handle subject-specific weighted averages of the z-specific individual controlled direct effects with respect to a user-supplied conditional distribution $Q_0(\cdot \mid W)$. Therefore, for the remainder of the paper we focus on estimation of the following parameter of interest:

$$DE(a, V) = DE(a, V \mid Q_0) \equiv E\{\sum_z (Y_{az} - Y_{0z}) Q_0(z \mid W) \mid V\}, \qquad (9)$$

where Q_0 could be known, or it could be the unknown $P(Z_0 = z \mid W) = P(Z = z \mid A = 0, W)$ in which case this definition reduces to (8). Geneletti (2007) and Didelez et al. (2006) make a related point (in a non-counterfactual framework) in their work on standardized direct effects.

2.3 Model for the direct effect parameter.

If $DE(a, V)$ is high-dimensional (e.g., A or V is continuous and/or has multiple dimensions), it is sensible practice to model this function. Consider a user-supplied model $\beta \to m(a, V \mid \beta)$ for this direct effect parameter $DE(a, V)$ in terms of a Euclidean parameter β:

$$DE(a, V) = E\{\sum_z (Y_{az} - Y_{0z}) Q_0(z \mid W) \mid V\} = m(a, V \mid \beta_0). \qquad (10)$$

This parametrization must be chosen so that it satisfies $m(0, V \mid \beta) = 0$ for all V and β. The true β_0 now represents the parameter of interest of the true data generating distribution P_0.

2.4 Models for the observed data distribution.

We note that, if Q_0 is a known conditional distribution, then $DE(a, V)$ is a parameter of the distribution of the full data structure

$$X^* \equiv [\{Y_{az} : (a, z) \in \mathcal{B}\}, W],$$

and (A, Z) can now be viewed as the joint missingness variable defining the observed data structure

$$O = (W, A, Z, Y_{AZ}), \tag{11}$$

where we assume that (A, Z) is randomized as defined by (2), or equivalently, that this joint missingness mechanism satisfies coarsening at random (van der Laan and Robins (2002a)). The variable (A, Z) is referred to as a "joint missingness variable" because, while the full data consist of counterfactuals indexed by each possible value for this joint variable, the observed data consist only of those counterfactuals indexed by the observed values of A and Z. Thus the observed value of (A, Z) determines which counterfactuals are observed and which are missing. We will denote the missing data model for P_0, defined by (11), (2), and (10), with $\mathcal{M}^*(CAR)$.

On the other hand, if $Q_0 = P(Z_0 \mid W)$, then β_0 is a parameter of the distribution of the full data structure

$$X \equiv [(Z_a : a \in \mathcal{A}), \{Y_{az} : (a, z) \in \mathcal{B}\}, W],$$

where now only A plays the missingness variable. In this case, in order to identify Q_0 one will also need the randomization assumption (5), and thus the model for P_0 is defined by (1), (2), (5), and (10). We will denote this latter model for P_0 with $\mathcal{M}(CAR)$. Our approach for construction of estimating functions for β_0 will be based on the missing data model $\mathcal{M}^*(CAR)$ for X^* assuming Q_0 is known. Simple substitution of estimators of Q_0 now also results in the wished class of estimators of β_0 in the model $\mathcal{M}(CAR)$.

3 Estimation.

In this section we present the IPCW estimating functions and corresponding estimators. Subsequently, we present the general class of DR-IPCW estimating functions, and corresponding DR-IPCW estimators. We discuss implementation, including the estimation of nuisance parameters. We then present a likelihood-based estimator that uses the identifiability result (7) rather than a model for the direct effect. Finally, we compare the properties of the three classes of estimators.

DOI: 10.2202/1557-4679.1064

3.1 Inverse probability of censoring weighted estimating functions.

Let $g_0(\cdot \mid X^*)$ denote the true conditional probability distribution of (A, Z) given X^*, and let g denote elements of our model for this conditional distribution. We refer to g as the censoring mechanism, since it determines which elements of the full data are seen and which are missing. The estimating functions developed in this section involve weighting by the inverse of g, and so are termed "inverse probability of censoring weighted" (IPCW) functions. Note that $g_0(A, Z \mid X^*) = g_0(A \mid W)g_0(Z \mid A, W)$ (by 2). Recall that $Q_0(\cdot \mid W)$ is either user-supplied and known, or equals the unknown $P_{Z_0|W} = P_{Z|A=0,W}$ (by 5).

Consider the following class of IPCW estimating functions for β_0, indexed by user-supplied functions

$$h(A, V) = \{h_1(A, V), h_2(V), g^*(A \mid V)\}$$

of A, V and nuisance parameter g:

$$D_{h,IPCW}(O \mid \beta, g, Q_0) \equiv \frac{g^*(A \mid V)}{g(A, Z \mid X^*)} \times \tag{12}$$
$$[h_1(A, V) - E_{g^*}\{h_1(A, V) \mid V\}]Q_0(Z \mid W)\{Y - m(A, V \mid \beta) - h_2(V)\}.$$

Here and subsequently we assume that $h(A, V)$ is a uniformly bounded function, so that the expectation and variance of all resulting estimating functions are well defined. Within that constraint, $h_1(A, V)$ can be any function of A and V, h_2 can be any function of V, and $g^*(\cdot \mid V)$ can be any conditional density of A, given V. A recommended choice for $h(A, V)$ will be given in Section 3.2. However, as is shown below, choice of $h(A, V)$ will not affect estimator consistency, therefore we refer to choice of a function $h(A, V)$ as an index for the IPCW estimating function, rather than as a nuisance parameter.

We make the following assumption regarding the censoring mechanism on the joint missingness variable (A, Z):

$$\max_{(a,z)\in\mathcal{B}} \frac{h_1(a, V)}{g_0(a, z \mid W)} < \infty \text{ a.e..} \tag{13}$$

For (13) to hold, it is sufficient that each possible combination of the treatment A and the level of intermediate Z occurs with some positive probability, regardless of baseline covariates W. In other words, baseline covariates should

not deterministically predict treatment, and treatment and baseline covariates should not deterministically predict the intermediate.

The following lemma establishes that, under (13), the IPCW estimating functions are indeed unbiased for β_0 at a correctly specified censoring mechanism g_0.

Lemma 1 *Assume model $\mathcal{M}^*(CAR)$ for P_0 and assumption (13).*
Then for any function $h = (h_1, h_2, g^*)$ of A, V

$$E_{P_0} D_{h,IPCW}(O \mid \beta_0, g_0, Q_0) = 0.$$

Proof of Lemma 1.

For notational convenience, we suppress the "IPCW" labelling. Firstly, we condition on $X^* = \{(Y_{az} : a, z), W\}$, which corresponds to integrating over A, Z w.r.t. $g_0(A, Z \mid W)$. This yields

$$\begin{aligned}
&E\{D_h(O \mid \beta_0, g_0, Q_0) \mid X^*\} \\
&= \textstyle\sum_{a,z} g^*(a \mid V)[h_1(a, V) - E_{g^*}\{h_1(A, V) \mid V\}]\times \\
&Q_0(z \mid W)\{Y_{az} - m(a, V \mid \beta_0) - h_2(V)\} \\
&= \textstyle\sum_a g^*(a \mid V)[h_1(a, V) - E_{g^*}\{h_1(A, V) \mid V\}]\times \\
&\{\textstyle\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) + \sum_z Y_{0z} Q_0(z \mid W) - m(a, V \mid \beta_0) - h_2(V)\}.
\end{aligned}$$

At the first equality, we relied on (13) so that the denominator $g_0(a, z \mid X^*)$ cancels out for all a, z for which $h_1(a, V) \neq 0$. Conditioning on V now yields,

$$\begin{aligned}
&\sum_a g^*(a \mid V)[h_1(a, V) - E_{g^*}\{h_1(A, V) \mid V\}] \times \\
&\qquad \left[E\{\sum_z Y_{0z} Q_0(z \mid W) \mid V\} - h_2(V)\right] \\
&\equiv \sum_a g^*(a \mid V)[h_1(a, V) - E_{g^*}\{h_1(A, V) \mid V\}] \times h_2^*(V),
\end{aligned}$$

where $h_2^*(V)$ is defined as $E\{\sum_z Y_{0z} Q_0(z \mid W) \mid V\} - h_2(V)$.

The $\sum_a g^*(a \mid V)[h_1(a, V) - E_{g^*}\{h_1(A, V) \mid V\}] = 0$, which completes the proof of Lemma 1. □

3.2 Inverse probability of censoring weighted estimators.

Let g_n be an estimator of the censoring mechanism $g_0(A, Z \mid W) = g_0(A \mid W) g_0(Z \mid A, W)$. If the weight function Q_0 is unknown, then let Q_{0n} be an

estimator of Q_0, but otherwise $Q_{0n} = Q_0$. The corresponding IPCW estimator of β_0 is now defined as the solution $\beta_{n,IPCW}$ of the estimating equation in β:

$$0 = \sum_{i=1}^{n} D_{h_n,IPCW}(O_i \mid \beta, g_n, Q_{0n}),$$

Where we assume the existence and uniquesness of a solution to the estimating equation, which is not obvious in the case that $m(A, V \mid \beta)$ is non-linear.

3.3 Double robust IPCW estimating functions.

van der Laan and Robins (2002a) (Theorem 1.3) show that one can orthogonalize an IPCW estimating function with respect to nuisance parameters. This procedure results in estimating functions which are more efficient than the initial IPCW estimating function, and in addition are maximally robust to nuisance parameter misspecification.

This orthogonalization is achieved by subtracting from the IPCW estimating function its projection on T_{CAR}, where $T_{CAR} \subset L_0^2(P_0)$ equals all possible nuisance scores corresponding with one dimensional fluctuations of the true missingness mechanism $g_0(A, Z \mid X^*)$ only assuming CAR (i.e.,$g_0(A, Z \mid X^*) = g_0(A, Z \mid W)$). That is, one subtracts from the IPCW estimating function its projection on the sub-Hilbert space of functions of A, Z, W with conditional mean zero, given W, within the Hilbert space $L_0^2(P_0)$ of functions of the observed data structure O with mean zero and finite variance, endowed with inner product $\langle f_1, f_2 \rangle_{P_0} \equiv E_{P_0} f_1(O) f_2(O)$ being the covariance operator. Thus these orthogonalized estimating functions are derived as $D_{h,DR} = D_{h,IPCW}(O) - \{E(D_{h,IPCW}(O) \mid A, Z, W) - E(D_{h,IPCW}(O) \mid W)\}$.

In the case of our IPCW estimating function for $DE(a, V)$ (13),we have that

$$\begin{aligned} &E\{D_{h,IPCW}(O \mid \beta, g, Q_0) \mid A, Z, W\} = \\ &\tfrac{g^*(A|V)}{g(A,Z|X^*)}[h_1(A, V) - E_{g^*}\{h_1(A, V) \mid V\}]Q_0(Z \mid W)\times \\ &\{E(Y \mid A, Z, W) - m(A, V \mid \beta) - h_2(V)\}. \end{aligned}$$

Thus,

$$\begin{aligned} &E\{D_{h,IPCW}(O \mid \beta, g, Q_0) \mid W\} = \\ &\textstyle\sum_{a,z} g^*(a \mid V)[h_1(a, V) - E_{g^*}\{h_1(A, V) \mid V\}]Q_0(z \mid W)\times \\ &\{E(Y \mid A = a, Z = z, W) - m(a, V \mid \beta) - h_2(V)\}. \end{aligned}$$

If we let $Q_Y(A, Z, W)$ represent a parameter value for $Q_{Y0}(A, Z, W) = E_{P_0}(Y \mid$

A, Z, W), then we have the following double robust estimating function:

$$D_{h,DR}(O \mid \beta, g, Q_Y, Q_0) = \\ \frac{g^*(A|V)}{g(A,Z|X^*)}[h_1(A,V) - E_{g^*}\{h_1(A,V) \mid V\}]Q_0(Z \mid W)\{Y - Q_Y(A,Z,W)\} \\ + \sum_{a,z} g^*(a \mid V)[h_1(a,V) - E_{g^*}\{h_1(A,V) \mid V\}]Q_0(z \mid W)\times \\ \{Q_Y(a,z,W) - m(a,V \mid \beta) - h_2(V)\}.$$

These estimating functions are double robust w.r.t. the pair of nuisance parameters (g_0, Q_{Y0}), where the term double robust refers to the fact that the estimator retains its consistency if at least one of this pair of nuisance parameters is consistently estimated.

Result 1 *Consider the class of double robust IPCW estimating functions:*

$$\{(O, \beta, g, Q_Y) \rightarrow D_{h,DR}(O \mid \beta, g, Q_Y, Q_0) : h\}.$$

If (13) holds at g, or in other words,

$$\max_{(a,z)\in\mathcal{B}} \frac{h_1(a,V)}{g(a,z \mid W)} < \infty \ a.e., \tag{14}$$

then for any index h

$$E_{P_0} D_{h,DR}(O \mid \beta_0, g_n, Q_{Y_n}, Q_0) = 0 \text{ if either } g = g_0 \text{ or } Q_Y = Q_{Y0}. \tag{15}$$

This is formally proved by application of the general estimating function theory (Section 1.6, van der Laan and Robins (2002a)), and is straightforward to verify explicitly. That is, this estimating function for β_0, which is indexed by two nuisance parameters, is unbiased if (13) holds at the possibly misspecified missingness-mechanism g, and one of the two nuisance parameters $g_0 = p_{A,Z|W}$, $Q_{Y0} = E_{P_0}(Y \mid A, Z, W)$ is correctly specified as well.

3.4 Double robust IPCW estimators.

Let h_n, g_n, Q_{Yn} be estimators of h^*, g_0, Q_{Y0}. The corresponding DR-IPCW-estimator of β_0 is now defined as the solution $\beta_{n,DR}$ of the estimating equation in β:

$$0 = \sum_{i=1}^{n} D_{h_n}(O_i \mid \beta, g_n, Q_{Yn}, Q_0).$$

If the weight function Q_0 in the definition of β_0 is unknown, then one replaces Q_0 by an estimator Q_{0n}.

DOI: 10.2202/1557-4679.1064

3.5 Implementation.

Choice of an index.

One possible choice $h^*(A,V)$ for the function $h(A,V)$ indexing the IPCW-estimating functions $D_{h,IPCW}$ is given by choosing

$$\begin{aligned} h_1(A,V) &= \frac{d}{d\beta_0} m(A,V \mid \beta_0) \\ h_2(V) &= m_0(V) = E_{P_0}\{\sum_z Y_{0z} Q_0(z \mid W) \mid V\} \\ g^*(A \mid V) &= g_0(A \mid V). \end{aligned}$$

We note that this choice is based on the known optimal choice of an analagous index function for the total causal effect parameter estimated in point treatment structural nested mean model (van der Laan and Robins, 2002a). The current model, however, is clearly distinct and further work is needed to determine optimal (maximally efficient) choices of $h(A,V)$.

Let h_n be an estimator of the function h. In the case that $h(A,V)$ is chosen to be $h^*(A,V)$, h_n can be estimated based on substitution of an estimator g_n^* of $g_0 = p_{A|V}$, and a regression estimator $h_{2n}(V)$ of m_0. Since

$$m_0(V) = E_{P_0}\{\sum_z Q_0(z \mid W) E_{P_0}(Y \mid A=0, Z=z, W) \mid V\},$$

we have that h_{2n} can in this case be obtained by regressing $\sum_z Q_0(z \mid W_i)\hat{E}(Y \mid A=0, Z=z, W_i)$ on V_i according to a working model $m(V \mid \eta)$ for $m_0(V)$, where $i = 1, ..., n$ is an index for the experimental unit. Note, however, that it is perfectly acceptable to just choose $h_2(V) = 0$. If $m(A,V \mid \beta)$ is non-linear in β, then $h_1(A,V)$ depends on β_0, so that one will need an initial estimator of β_0 to estimate h_1, which can be obtained by first using an h_1 at an initial guess of β.

Nuisance parameter estimation.

The estimating functions presented for β_0 in model $\mathcal{M}^*(CAR)$ depend on potentially high-dimensional nuisance parameters, and thus construction of corresponding estimators for the direct effect parameter requires specification of nuisance parameter models. In order to address the curse of dimensionality in the model $\mathcal{M}^*(CAR)$, one needs to assume a model for at least one of the following two nuisance parameters:

1. $g_0(\cdot \mid X^*) = g_0(Z|A,W)g_0(A|W)$

2. $Q_{Y0} = E_{P_0}(Y|A, Z, W)$.

Given user-supplied models for these nuisance parameters, they can be estimated using maximum likelihood. In other words, the nuisance parameter estimators can be defined as the maximizers over user-supplied models of the relevant partial likelihoods given by

$$\begin{aligned} L(f_{A|W}) &= \prod_{i=1}^{n} f_{A|W}(A_i \mid W_i) \\ L(f_{Z|A,W}) &= \prod_{i=1}^{n} f_{Z|A,W}(Z_i \mid A_i, W_i) \\ L(f_{Y|A,Z,W}) &= \prod_{i=1}^{n} f_{Y|A,Z,W}(Y_i \mid A_i, Z_i, W_i), \end{aligned}$$

where $i = 1, ..., n$ is an index for the experimental unit. Other (e.g., estimating-function based) procedures for estimation of the nuisance parameters can be used as well. In particular, one can use cross-validation methodology to data-adaptively select the models for these parameters.

Table 1 summarizes the contributions of nuisance parameters to the estimators presented. Note that, in the case that $Q_0 = P_{Z_0|W}$ and is thus unknown, Q_0 serves as an additional nuisance parameter, which, again, can be estimated by maximizing the likelihood of a user-supplied model, or via estimating function-based procedures. (Recall that under (5) $P_{Z_0|W} = P_{Z|A=0,W}$.)

Implementation.

If $m(\cdot \mid \beta)$ is linear in β, then the estimating equations presented are just a linear system of equations in β, and can thus be solved trivially in closed form. For general parameterizations the estimators can be computed using the Newton-Raphson algorithm with a standard line search correction guaranteeing that at each step the Euclidean norm of the estimating equation decreases (to zero). In this case one can use $\beta_{n,IPCW}$ as initial estimator. For more details, we refer to van der Laan and Robins (2002b).

We further note that if one chooses $h_2(V) = m_0(V)$ and is willing to assume a correctly specified model for m_0, a class of IPCW estimators can be represented as a weighted least squares estimator and implemented using standard software van der Laan and Petersen (2005).

DOI: 10.2202/1557-4679.1064

Table 1: Comparison of inverse probability of censoring weighted, double robust inverse probability of censoring weighted, and likelihood-based direct effect estimators with respect to dependence on nuisance parameters. *IPCW=inverse probability of censoring weighted, DR-IPCW=double robust inverse probability of censoring weighted.* †: *When* $Q_0 = P_{Z_0|W}$. ‡: *Either* g_0 *or* Q_{Y0} *must be correctly specified.*

Used in estimator.			
Estimator	$g_0(\cdot\|X^*)$	$Q_{Y0} = E_{P_0}(Y\|A,Z,W)$	Q_0
IPCW	Yes	No	Yes
DR-IPCW	Yes	Yes	Yes
Likelihood-based	No	Yes	Yes
Consistent estimation required for consistency of estimator.			
Estimator	$g_0(\cdot\|X^*)$	$Q_{Y0} = E_{P_0}(Y\|A,Z,W)$	Q_0
IPCW	Yes	No	Yes†
DR-IPCW	No‡	No‡	Yes†
Likelihood-based	No	Yes	Yes†

3.6 Likelihood-based estimator.

The preceding sections have presented two estimators (IPCW and DR-IPCW) based on assuming a model for the direct effect parameter of interest $DE(a,V)$. For comparison, we review an alternative approach based on the identifiability results of Robins and Greenland (1992), Pearl (2000), van der Laan and Petersen (2004) and Petersen et al. (2006).

Consider the identifiability result (7) for

$$DE(a,W) = E\{\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) \mid W\}$$

given by:

$$DE(a,W) = \int_z \{E(Y \mid A=a, Z=z, W) - E(Y \mid A=0, Z=z, W)\} Q_0(z \mid W). \tag{16}$$

This identifiability result suggests the following likelihood-based estimator, based on noting that $E_{P_0}\{DE(a,W) \mid V\} = m(a,V \mid \beta_0)$:

1. Estimate $Q_{Y0}(A,Z,W) = E(Y \mid A,Z,W)$ as Q_{Yn}.

2. Plug the estimator Q_{Yn} into the identifiability result (17) to obtain a fitted $\hat{DE}(a, W)$. In the case that $Q_0 = P_{Z_0|W}$, this will also require estimating Q_0 as $Q_n = \hat{E}(Z|A = 0, W)$.

3. Regress the vector $\{\hat{DE}(a, W_i) : a\}$, consisting of a fitted W-specific direct effect estimate for each subject i and possible treatment value a, on V according to the model $E\{\hat{DE}(a, W) \mid V\} = m(a, V \mid \beta)$.

Alternative methods for obtaining such a substitution-type estimator of $DE(a, W)$ are discussed in detail in (van der Laan and Petersen, 2004).

3.7 Discussion of the three types of estimators.

Table 1 summarizes how the consistency of each direct effect estimator (IPCW, DR-IPCW, and likelihood-based) depends on consistent nuisance parameter estimation. In the case that $Q_0 = P_{Z_0|W}$, and is thus unknown, the consistency of all three classes of estimators also relies on consistent estimation of $P_{Z|A,W}$.

We note that the consistency of our estimators of β_0 does not depend on choice of an index function $h(A, V)$, or consistent estimation of its components. In particular, given the choice $h_2(V) = m_0(V)$, the validity of the working model for $m_0(V) \equiv E\{\sum_z Y_{0z} Q_0(z \mid W) \mid V\}$ (i.e., the consistency of the corresponding estimator of m_0) only potentially affects the efficiency of our estimators of β_0; it does *not* affect the consistency and asymptotic linearity of these estimators.

The consistency of the IPCW-estimator of β_0 relies on the consistency of g_n as an estimator of the censoring mechanism $g_0(\cdot \mid X^*)$, and on (13). The consistency of the likelihood-based estimator of β_0 relies on the consistency of Q_{Yn} as estimator of $E(Y \mid A, Z, W)$. Finally, if (13) holds for g_0, then the consistency of the DR-estimator of β_0 relies on the consistent estimation of either g_0 or Q_{Y0}, but, if (13) fails to hold for g_0, then it fully relies on consistent estimation of Q_{Y0}. That is, the DR-IPCW estimator remains consistent when (13) is violated as long as Q_{Y0} is estimated consistently and (13) holds under the (incorrectly specified) estimate g_n employed in the estimator. The DR-IPCW estimator thus has the attractive property of being the most non-parametric estimator, retaining its consistency if either the IPCW-estimator or the likelihood-based estimator is consistent

Note that violation of (13) not only implies that the IPCW estimator is inconsistent, it further reflects a lack of support in the data for the parameter β_0. Consider the case in which both treatment A and intermediate Z are binary. If subjects in a given covariate stratum $W = w$ never experienced, for example, both the presence of treatment and the presence of the

intermediate ($(A = 1, Z = 1)$), no information exists in the data about the outcomes under this combination of exposures. Thus both the DR-IPCW and Likelihood-based approaches will rely on extrapolation from other areas of the data (for example, from other covariate strata where this combination of treatment and intermediate was observed). For this reason, we consider it advisable to investigate the degree of bias in the IPCW estimator due to violation of (13), regardless if one is also employing the DR-IPCW or likelihood based estimators. In the case that a violation of (13) is estimated to result in considerable bias to the IPCW estimator, then the likelihood-based estimator might be the preferred estimator, since in this case the DR-IPCW estimator is no more robust than the likelihood-based estimator; both now fully rely on the consistent estimation of $E(Y \mid A, Z, W)$.

Wang et al. (2006) suggest one approach to estimating the bias to the IPCW estimator caused by the violation of (13). Briefly, the sampling distribution of the IPCW estimator is computed under a maximum likelihood estimator of the data-generating distribution (i.e., a parametric bootstrap is implemented), and the mean of the sampling distribution is compared with β_0 as calculated from the fitted likelihood.

4 Double robust targeted MLE.

Above we discussed estimators as solutions of the double robust IPCW estimating equations. An important issue with defining an estimator of β_0 as a solution of estimating equations is that such solutions are typically not compatible with a particular probability distribution of the data. In addition, an estimating equation may have multiple solutions, and this approach is lacking in criteria by which to select among a set of possible solutions. Targeted maximum likelihood estimation (van der Laan and Rubin (1996)) address these issues by providing a maximum likelihood-based estimation procedure resulting in an estimator $Q_n = (Q_{Wn}, Q_{Yn})$ where Q_{Wn} represents an estimator of the distribution of W, and Q_{Yn} represents an estimator of $E(Y|A, W)$. The corresponding substitution estimator $\beta_n = \beta(Q_n)$ then solves the wished double robust estimating equation.

We start with noting that the double robust estimating functions indexed

by h can be decomposed as a sum of two estimating functions:

$$
\begin{aligned}
&D^*_{h,DR}(O \mid \beta_0, g_0, Q_{Y0}) = \\
&\tfrac{g^*(A|V)}{g_0(A,Z|X^*)} \{h_1(A,V) - E_{g^*}(h_1(A,V) \mid V)\} Q_0(Z \mid W)(Y - Q_{Y0}(A,Z,W)) \\
&+ \textstyle\sum_{a,z} g^*(a \mid V) \{h_1(a,V) - E_{g^*}(h_1(A,V) \mid V)\} Q_0(z \mid W) \times \\
&\{Q_{Y0}(a,z,W) - m(a,V \mid \beta_0) - h_2(V)\} \\
&\equiv D^*_{1h}(g_0, Q_{Y0}) + D^*_{2h}(\beta_0, Q_{Y0}),
\end{aligned}
$$

where we suppress the dependence on $Q_0(Z \mid W)$.

This class of estimating functions indexed by h are double robust in the sense that they are solved at the true β_0 if either $g = g_0$ or $Q = Q_0$, and the model is correctly specified.

Based on efficiency considerations, as target choice $h = (g^*, h_1, h_2)$ we recommended $h_1(a,V) = \frac{d}{d\beta_0} m(a,V \mid \beta_0)$, $h_2(V) = m_0(V) \equiv E(\sum_z Q_0(z \mid W) Q_{Y0}(0,z,W) \mid V)$, and $g^*(A \mid V) = g_0(A \mid V)$. If $m(\cdot \mid \beta)$ is linear in β, then h_1 is known. Either way, we replace each of these target choices by estimates of the corresponding quantities, resulting in a choice h_n with an asymptotic limit h_∞, not necessarily equal to this wished choice h. We note that under the assumption that $m(a,V \mid \beta_0)$ is correctly specified, $E(\sum_z Y_{az} Q_0(z \mid W) \mid V) = m(a,V \mid \beta_0) + m_0(V)$ so that one can view $m(a,V \mid \beta_0) + m_0(V)$ as a semi-parametric additive causal regression model modeling direct effects.

Since it is hard to construct an estimator Q_{Yn} of Q_{Y0} satisfying a particular direct effect model $m(\cdot \mid \beta)$, we wish to work in the nonparametric model in which $m(\cdot \mid \beta)$ is merely viewed as a working model. In this nonparametric model one can view the (standardized version of the) recommended choice of double robust estimating function as an efficient influence curve in a nonparametric model for a parameter $Q = (Q_W, Q_Y) \rightarrow \beta(Q)$ defined nonparametrically as a solution of the second component equation: $\beta(Q)$ is a β solving $E_{Q_W} D_{2h^*}(\beta, Q_Y) = 0$, where Q_W denotes a distribution of W and E_{Q_W} denotes the expectation with respect to the distribution Q_W.

Our particular choice of mapping from a distribution Q_W of W and the conditional mean Q_Y into $\beta(Q_W, Q_Y)$ solving this equation is now defined. We note that if the direct effect model $m(\cdot \mid \beta)$ is correctly specified, then $\beta(Q_0)$ corresponds with the true parameter in this direct effect model.

Nonparametric definition of $\beta(Q)$ solving the second component of the efficient influence curve equation:
We note that, given an estimator $Q = (Q_W, Q_Y)$, we can define $(\beta(Q), m(Q))$ as follows. We consider the case that the marginal distribution of W under Q is the empirical probability distribution of W.

DOI: 10.2202/1557-4679.1064

This definition relies on the specification of an initial estimator β^0, m^0 (which could be implied by $Q = (Q_W, Q_Y)$ itself) of β_0, m_0. Firstly, we compute the following weighted linear regression estimator:

$$\epsilon_n^0 \equiv \arg\max_{\epsilon} \sum_i \sum_{a,z} g_n^*(a \mid V_i) Q_0(z \mid W)$$
$$\left\{ Q_Y(a, z, W_i) - m_{\beta^0+\epsilon}(a, V_i) - m^0(V_i) - \epsilon h_n^*(V_i) \right\}^2,$$

where we define $h_n^*(V) = -E_{g_n^*}(h_1(A, V) \mid V) = -E_{g_n^*}(\frac{d}{d\beta^0} m_{\beta^0}(A, V) \mid V)$.

Thus, for example, if $m(a, V \mid \beta) = \beta aV$ is linear, then ϵ_n^0 is obtained as a repeated measures weighted linear least squares regression fit with off-set $\beta^0 aV + m^0(V)$, adding covariate extension $\epsilon\{aV + h^*(V)\}$, where the weights are $g*(a \mid V_i)Q_0(z \mid W_i)$.

Let $\beta^1 = \beta^0 + \epsilon_n^0$ and $m^1 = m^0 + \epsilon h^*$ be the updates. These updates can be interpreted as an estimator of β_0, m_0 which, given an initial estimator (β^0, m^0), specifically targets β_0.

In general, if one wishes to also update h^* itself based on the newly obtained estimate β^1, then, we can iterate this process till convergence (i.e., $\epsilon_n^k \approx 0$) and let $\beta(Q) = \beta^k$ and $m(Q) = m^k$ for this large enough choice k. If either m_β is linear or one simply uses a fixed h^* according to its initial estimate, then the convergence occurs in a single step so that $\beta_n^k = \beta^1$ and $m_n^k = m^1$. For simplicity, and since it comes without cost of asymptotic performance, we recommend to use a fixed h^* so that the first step β^1 already represents the evaluation $\beta(Q)$.

Because the score of the used ϵ-extension for the final k is solved at $\epsilon = 0$ it follows that $\beta(Q), m(Q)$ solves the following estimating equation:

$$\begin{aligned} 0 \;=\; & \sum_i \sum_{a,z} g_n^*(a \mid V_i) Q_0(z \mid W_i) \left\{ \frac{d}{d\beta_n^0} m_{\beta_n^0}(a, V_i) + h_n^*(V_i) \right\} \times \\ & (Q^0(a, z, W_i) - m_{\beta(Q)}(a, V_i) - m(Q)(V_i)), \end{aligned}$$

or equivalently $\sum_i D_{2h_n}(\beta(Q_W, Q_Y), Q_Y)(O_i) = 0$ with $h_n = (g_n^*, h_n^*, h_2 = m(Q))$.

Solving the first component of the efficient influence curve equation using targeted MLE:
Let $Q^0 = (Q_Y^0, Q_W^0)$ be an initial estimator estimating $Q_{Y0}(A, Z, W) = E_0(Y \mid A, Z, W)$ according to a regression model, and estimating the marginal distribution of W with the empirical distribution of W. Targeted maximum likelihood estimation involves updating such an initial estimator by maximizing the likelihood in a particular direction targeting the parameter $\beta(Q)$ of

interest, thereby reducing the bias of this initial likelihood based estimator $\beta(Q^0)$.

Let $Q^0(\epsilon)$ be obtained by adding $\epsilon C_n(A,Z,W)$ to the regression fit $Q_Y^0(A,W)$, while keeping Q_W^0 the same, where

$$C_n(A,Z,W) = \frac{g_n^*(A \mid V)}{g_n(A,Z \mid X^*)} Q_0(Z \mid W) \left\{ \frac{d}{d\beta^0} m_{\beta^0}(A,V) + h_n^*(V) \right\}.$$

Let ϵ_n^0 be the MLE according to a normal regression model:

$$\epsilon_n^0 = \arg\min_{\epsilon} \sum_{i=1}^{n} (Y_i - Q_Y(\epsilon)(A_i,Z_i,W_i))^2.$$

Let Q_n^1 be the corresponding update. In principle, if one decides to update C_n at each step, we iterate this till $\epsilon_n^k \approx 0$ and thereby Q_n^k solves

$$0 = \sum_i D_{1h_n}(Q_n^k, g_n)(O_i).$$

However, if the covariate C_n does not depend on Q_n^k (i.e., we fix it at its initial estimate), then it follows that convergence occurs at the first step so that we already have

$$0 = \sum_i D_{1h_n}(Q_n^1, g_n)(O_i)$$

for $h_n = (g_n^*, h_n^*, h_2)$ for arbitrary choice h_2 (since D_{1h} does not depend on h_2). Again, we recommend to use a fixed C_n so that the targeted MLE is attained in a single step.

The targeted MLE solves the efficient influence curve equation: We start out with specifying an estimator g_n^*, corresponding estimate h_n^*, and an estimator $h_{2n}(V) = m_n^0(V)$ of $\theta_0(V)$, resulting in $h_n = (g_n^*, h_n^*, h_{2n})$. In addition, we specify an initial estimator $Q_n^0 = (Q_{Wn}^0, Q_{Yn}^0)$ of (Q_{W0}, Q_{Y0}), where we set Q_{Wn}^0 equal to the empirical distribution of $W_1, \ldots, W_n$. Given Q_n^0, we find the one step targeted MLE $Q_n^1 = Q_n^0(\epsilon_n^0)$ (defined above) solving

$$0 = \sum_i D_{1,g_n^*,h_n^*}(Q_n^1, g_n)(O_i) = 0.$$

Given Q_n^1 (and say corresponding β_n^0, m_n^0), we determine $\beta_n^1 = \beta(Q_n^1), \theta_n^1 = \theta(Q_n^1)$ defined above as a targeted semi-parametric repeated measures regression solving

$$0 = \sum_i D_{2,g_n^*,h_n^*,m(Q_n^1)}(\beta^1, Q_n^1)(O_i) = 0.$$

We conclude that the targeted MLE $\beta_n^1 = \beta(Q_n^1)$ solves the double robust estimating equation

$$0 = \sum_i D_{h_n,DR}(\beta^1 = \beta(Q_n^1), Q_n^1, g_n)(O_i) = 0,$$

where $h_n = (g_n^*, h_n^*, \theta(Q_n^1))$.

Since β_n^1 solves the double robust estimating equation, statistical inference for β_0 based on the targeted MLE proceeds in the same manner as for the DR-IPCW estimator, described in the Section which follows.

5 Statistical inference for direct effect models.

Consider the case that Q_0 is known. Under the above stated assumptions regarding correct estimation of the nuisance parameters g_0, Q_{Y0}, as well as regularity conditions guaranteeing that the second order terms of differences between g_n and g_0 and Q_{Yn} and Q_{Y0} are $o_P(1/\sqrt{n})$, it can be shown that the IPCW and DR-IPCW estimators β_n of β_0 are root-n consistent and that $\sqrt{n}(\beta_n - \beta_0)$ is asymptotically normally distributed with mean zero and certain variance (see Theorems 2.4 and 2.5 in van der Laan and Robins (2002a)). Under these regularity conditions one can also establish the asymptotic validity of the bootstrap for obtaining confidence regions for β_0. The resulting confidence regions can be used for testing purposes.

We focus here on the application of Theorem 2.4 in van der Laan and Robins (2002a), which relies on assuming a correctly specified model for g_0, to statistical inference for the direct effect parameter $DE(a, V)$. We limit our discussion to the DR-IPCW estimator of β_0, which we refer to in this section as β_n, since the IPCW estimating functions presented in Section 3.1 simply correspond with setting $Q_Y = 0$ in the double robust estimating functions.

First, in the unrealistic case that $g_n = g_0$, then under regularity conditions specified in Theorem 2.4, we have that

$$\beta_n - \beta_0 = \frac{1}{n}\sum_{i=1}^n IC_0(O_i) + o_P(1/\sqrt{n}),$$

where the influence curve IC_0 is given by

$$IC_0(O) = -c(\beta_0)^{-1} D_{h,DR}(O \mid \beta_0, g_0, Q_{Y1}, Q_0)$$

and Q_{Y1} denotes the possibly misspecified limit of Q_{Yn} (e.g., $Q_{Yn} = 0 = Q_{Y1}$). Here $c(\beta_0) \equiv d/d\beta_0 E D_{h,DR}(O \mid \beta_0, g_0, Q_{Y1}, Q_0)$ is the matrix obtained by

differentiating the expectation of the estimating function w.r.t. β at β_0. The latter matrix can be easily estimated as the derivative matrix of the actual estimating equation.

If one now uses an actual maximum likelihood estimator g_n of the censoring mechanism g_0 according to a correctly specified model with tangent space $T_G(P_0) \subset T_{CAR}(P_0)$ (space spanned by nuisance scores for the true censoring mechanism g_0, where nuisance scores refer to the scores of one-dimensional sub-models, through the truth, varying only the nuisance parameter g_0), then Theorem 2.4 in van der Laan and Robins (2002a) states that the influence curve improves to

$$IC = IC_0 - \Pi(IC_0 \mid T_G(P_0)),$$

where $\Pi(IC_0 \mid T_G(P_0))$ denotes the projection of IC_0 onto $T_G(P_0)$ in the Hilbert space $L_0^2(P_0)$ (see also Theorem 2.3 van der Laan and Robins (2002a)). That is, subtracting the projection of IC_0 on the tangent space improves asymptotic efficiency.

In the special case that $Q_{Y1} = Q_{Y0}$ (that is, our regression estimator of $E(Y \mid A, Z, W)$ is asymptotically consistent), then IC_0 is already orthogonal to all possible censoring mechanism scores (i.e., $T_{CAR}(P_0)$) and the projection equals zero such that $IC = IC_0$. More generally, one can use IC_0 as a conservative influence curve. In other words, we can estimate the covariance matrix of β_n conservatively with

$$\Sigma_n \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{IC_0}(O_i) \hat{IC_0}(O_i)^\top,$$

where $\hat{IC_0}$ is an estimate of the true influence curve IC_0 obtained by substituting our estimators β_n, g_n, Q_{Yn} into the expression for IC_0:

$$\hat{IC_0}(O_i) = -c_n(\beta_n)^{-1} D_h(O_i \mid \beta_n, g_n, Q_{Yn}, Q_0),\ i = 1, \ldots, n.$$

One can now construct conservative confidence regions for β_0 based on the multivariate normal working model $\beta_n \sim N(\beta_0, \Sigma_n/n)$.

The advantage of the above conservative approach is that it requires no extra work beyond evaluation of the estimating equation (which was already needed in the construction of β_n), and it avoids computer intensive re-sampling.

We suggest that, even in the double robust model (Theorem 2.5 in van der Laan and Robins (2002a)) the above influence curve IC_0 will typically be conservative, but calculation of the true influence curve is now more involved. In general, we remind the reader that the bootstrap is asymptotically valid, and typically provides a more accurate estimate of the *finite* sample distribution

DOI: 10.2202/1557-4679.1064

than the influence curve (Wald-type) approach described above. Of note, while the non-parameteric bootstrap provides a straightforward approach to estimation of the variance, it will not serve to detect bias resulting from violation of (13). For this reason, as discussed above, we recommend complimentary use of a parametric bootstrap technique to investigate the presence of such violations, as discussed above.

In the case that Q_0 is unknown and thus estimated by an estimator Q_{0n}, then the influence curve of β_n equals the influence curve above, plus an additional component due to the estimation of Q_0, which can typically be calculated explicitly. If one wishes to avoid such calculations, we again suggest simply using the bootstrap.

6 Discussion.

The aim of this article is to give direct effect estimation the same place in the causal inference literature as estimation of total causal effects. Our approach has been to model the observed data distribution treating A, Z as a joint treatment or missingness variable, to only assume randomization (or no unmeasured confounders), and to define the direct effect parameter $DE(a, V)$ as the parameter of interest in this model. By adding the conditional independence assumption 6 and selecting as weight function $Q_0 = P_{Z_0|W}$, this parameter reduces to the conventional definition of a natural direct effect. However, if the conditional independence assumption fails to hold, $DE(a, V)$ remains a potentially interesting direct effect parameter, in that it can be interpreted as the population mean (within strata of V) of a subject-specific weighted average of the z-specific individual controlled direct effects, averaged with respect to a user-supplied conditional distribution $Q_0(\cdot|W)$.

A concern frequently raised in discussion of direct effects is that such effects do not correspond to a possible experiment. The concern with natural direct effects, in particular, arises because one of the counterfactuals used to define a natural direct effect, Y_{aZ_0}, will never be observed, as it corresponds to simultaneously setting treatment at level a, and also at level 0, the latter in order to observe Z_0. (Robins and Greenland (1992) note a possible exception in the context of randomized cross-over trials). Robins (2003) offers some perspective defending definitions of causal effects that lack correspondence to a possible experiment. Specifically, he defines manipulative causal DAG models as models in which causal effects could be checked (in principle) by manipulation/ experimental intervention, and shows that natural direct effects do not correspond to a manipulative DAG model. He goes on to defend the use of

non-manipulative DAG models with two major arguments. First, the reliance of such models on untestable assumptions should not be prohibitive, unless one is also willing to forgo estimation of total causal effects based on non-randomized data, which relies on similarly untestable assumptions regarding unmeasured confounding. Second, a major motivation for preferring causal effects that correspond to manipulative causal graphs is that in principle this allows assumptions/findings to be tested empirically. Based on this motivation, however, there is no reason that an effect which it is theoretically possible but practically impossible to reproduce using an experiment should be more appealing than an effect that is theoretically impossible to reproduce using an experiment; in neither case will checking assumptions be feasible. In addition, even a the hypothetical experiment corresponding to the causal effect is feasible, using such an experiment to actually check assumptions requires that the group on which we propose to intervene in order to check our assumptions is exchangeable with the group on which estimates are based; in practice such a group will rarely be available.

Petersen et al. (2006) emphasize that effects can be mechanistically interesting, even if they do not correspond to possible interventions. More importantly, they make the point that, ultimately, it is up to the researcher whether a counterfactual interpretation is meaningful for a given application. In the case that the researcher is not comfortable with such a counterfactual interpretation, or indeed, is philosophically uncomfortable considering "causal effects" in the absence of a corresponding experiment, then an alternative non-causal definition of direct effects can be employed. Namely, in a point treatment setting, controlled direct effects can be understood as the difference in outcome between individuals who are treated vs. untreated, and are exchangeable with respect to the values of all measured baseline covariates and the intermediate. A natural direct effect is then simply defined as a summary measure of the controlled direct effect, specifically a weighted average, where weighting is determined by distribution of intermediate given baseline covariates in absence of exposure.

Such a non-counterfactual definition can also be generalized to the case in which the treatment and or intermediate vary over time. Specifically, the controlled direct effect can be defined by first intervening on the likelihood by setting a and z without altering other equations or factors of the likelihood. The controlled direct effect can be estimated by evaluating the mean under this intervention likelihood, while the natural direct effect is based on averaging this difference in means with respect to a particular distribution of z. We note that these alternative definitions no longer depend on untestable assumptions regarding unmeasured confounding, nor are any additional assumptions

necessary to make the natural direct effect identifiable.

References

S.R. Cole and M.A. Hernan. Fallibility in estimating direct effects. *Epidemiology*, 31:163–165, 2002.

S. Deeks, J. Barbour, Martin J., Swanson M., and Grant R. Sustained CD4+ T cell response after virologic failure of protease inhibitor-based regimens in patients with human immunodeficiency virus infection. *J Infect Dis*, 181 (3):946–53, 2000.

V. Didelez, A.P. Dawid, and S. Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artifical Intelligence*, pages 138–146, 2006.

S. Geneletti. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society: Series B*, 69(2):199–215, 2007.

R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random: characterizations, conjectures and counter-examples. In D.Y. Lin and T.R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–94, New York, 1997. Springer Verlag.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *Annals of Statstics*, 19:2244–53, 1991.

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics*, 23:774–86, 1995.

J. Neyman. On the application of probability theory to agricultural experiments. *Statistical Science*, 5:465–480, 1990.

J. Pearl. Direct and indirect effects. In M. Kaufmann, editor, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. San Francisco, 2000.

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

C. Poole and J.S. Kaufman. What does standard adjustment for down- stream mediators tell us about social effect pathways. *American Journal of Epidemiology*, 151:s52, 2000.

J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J.M. Robins. Semantics of causal DAG models and the identification of direct and indirect effects. In P. Green, N. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 70–81. Oxford University Press, Oxford, 2003.

J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, pages 95–133. Springer, New York, 2000.

J.M. Robins. Association, causation, and marginal structural models. *Synthese*, 121:151–179, November 1999.

J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3:143–155, 1992.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–90, 1976.

D.B. Rubin. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics*, 6:34–58, 1978.

M.J. van der Laan and M.L. Petersen. Estmation of direct and indirect causal effects in longitudinal studies. Technical Report 155, Division of Biostatistics, University of California, Berkeley, 2004. URL `www.bepress.com/ucbbiostat/paper155/`.

M.J. van der Laan and M.L. Petersen. Direct effect models. Technical Report 187, Division of Biostatistics, University of California, Berkeley, 2005. URL `www.bepress.com/ucbbiostat/paper187/`.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2002a.

M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*, pages 118–119. Springer, New York, 2002b.

M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):11, 1996.

Y. Wang, M.L. Petersen, D.R. Bangsberg, and M.J. van der Laan. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. Technical Report 211, Division of Biostatistics, University of California, Berkeley, 2006. URL