University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 187

Direct Effect Models

Mark J. van der Laan* Maya L. Petersen†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, mayaliv@berkeley.edu

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 mediated by a given intermediate variable (the indirect effect of the treatment), and the component that is not mediated by that intermediate variable (the direct effect of the treatment) is often relevant to mechanistic understanding and to the design of clinical and public health interventions. Under the assumption of no-unmeasured confounders for treatment and the intermediate variable, Robins & Greenland (1992) define an individual direct effect as the counterfactual effect of a treatment on an outcome when the intermediate variable is set at the value it would have had if the individual had not been treated, and the population direct effect as the mean of these individual counterfactual direct effects. In this article we first generalize this definition of a direct effect. Given a user-supplied model for the population direct effect of treatment actions, possibly conditional on a user-supplied subset of the baseline co-variables, we propose inverse probability of treatment weighted estimators, likelihood-based estimators, and double robust inverse probability of treatment weighted estimators of the unknown parameters of this model. The inverse probability of treatment weighted estimator corresponds with a weighted regression and can thus be implemented with standard software.

1 Introduction and overview.

Estimation of the (total) causal effect of a treatment on an outcome is a primary focus of epidemiological and clinical research and has been the subject of major methodological advances. For an overview of various causal models and corresponding literature (counterfactual framework, marginal structural models, structural nested models, G-computation formula), and a presentation of a general locally efficient estimating function-based methodology for estimation of the corresponding causal parameters, we refer to van der Laan and Robins (2002). In many settings, it is of significant interest to identify the pathways by which a treatment is acting and to quantify the component of the treatment's effect that is and is not mediated by a given intermediate variable (the indirect and direct effects of the treatment, respectively). Estimation of the direct and indirect effects of a treatment can often inform mechanistic understanding of the treatment's action, as well as the design of clinical and public health interventions.

Robins and Greenland (1992) and Pearl (2000) have addressed the identification and estimation of direct and indirect causal effects in both the epidemiological and non-epidemiological literature. In order to define a direct causal effect, Robins and Greenland (1992) use the counterfactual framework, in which one assumes for a randomly sampled subject the existence of counterfactual outcomes Y_{az} and counterfactual intermediate variables Z_a under set values of treatment $A = a$ and the intermediate variable $Z = z$. The observed data can now be viewed as a missing data structure on these counterfactuals. Using this framework, an individual direct effect can be defined as $Y_{aZ_0} - Y_{0Z_0}$, or the counterfactual effect of a treatment $A = a$ on an outcome when the intermediate variable is set at the counterfactual value Z_0 that it would have had if the individual had not been treated (i.e, $A = 0$). Alternatively, an individual direct effect can be defined as $Y_{az} - Y_{0z}$, or the counterfactual effect of a treatment $A = a$ when the intermediate is set at a fixed level $Z = z$. In this article, we follow the lead of Pearl (2000) and call $Y_{aZ_0} - Y_{0Z_0}$ the *natural* individual direct effect, and $Y_{az} - Y_{0z}$ the *controlled* individual direct effect. The population direct effect is defined as the mean of these individual counterfactual direct effects.

In Robins and Greenland (1992), the authors discuss at length the limitations of multi-variable regression to estimate direct and indirect effects. Although frequently used, Robins and Greenland illustrate that this approach can lead to a biased estimate of the direct effect of treatment, even if there are no unmeasured confounders of the relationship between the treatment and outcome. In the setting of longitudinal treatment, unbiased estimation of direct effects using traditional multi-variable regression is often not possible. Robins and Greenland further introduce an additional assumption needed for natural direct effects to be identifiable from the observed data; Robins and Greenland (1992)'s No Interaction Assumption states that the individual direct effect at a fixed level of the intermediate variable (controlled direct effect) does not depend on the level at which the intermediate variable is fixed. Pearl (2000) provides an alternative identifying assumption for natural direct effects; he assumes that an individual's outcome under a fixed level of the intermediate variable does not depend on the level of the intermediate variable that the individual would have had under no treatment. Both of the assumptions proposed to date to make direct effects identifiable have been considered both unrealistic and restrictive for applications in epidemiology and clinical medicine.

In recent work (van der Laan and Petersen (2004)), we introduced an alternative assumption to make natural direct effects identifiable. We showed that the identifiability result of Pearl (2000) also holds under a new conditional independence assumption which states that, within strata of baseline co-variables, the individual direct causal effect at a fixed level of the intermediate variable (controlled direct effect) is independent (in the mean sense) of the no-treatment counterfactual intermediate variable. Using both theoretical arguments and an example drawn from our research, we argued that our assumption is more realistic and less restrictive than the assumptions of Robins and Greenland (1992) and Pearl (2000). We refer to Petersen et al. (2006) for an epidemiological discussion and interpretation of direct effects.

Under randomization assumptions (assumptions of no unmeasured confounders) and the conditional independence assumption, the natural direct effect parameter is identifiable from the observed data. In this article we first note that if the conditional independence assumption fails to hold, then the parameter that the identifiability result for the natural direct effect parameter $E(Y_{aZ_0} - Y_{0Z_0})$ targets equals the population mean of a subject-specific average of z-specific controlled direct effects, $Y_{az} - Y_{0z}$, w.r.t. to a conditional distribution of Z_0, given the subject's baseline co-variables. The latter is itself an interesting direct effect parameter, which can be interpreted as the population mean of a subject-specific average of its controlled individual direct effects. Therefore, in this article, we focus on modelling and estimation of this direct effect parameter, which happens to agree with the conventional natural direct effect parameter $E(Y_{aZ_0} - Y_{0Z_0})$ under our conditional independence assumption, but does not rely on it. In addition, we generalize this parameter by allowing the choice of conditional distribution used to obtain an average of the z-specific controlled individual direct effects $Y_{az} - Y_{0z}$ to be user-supplied.

Given a user-supplied parametrization (i.e., model) of this direct effect parameter, adjusted for a user-supplied subset of the baseline co-variables, in this article we are concerned with estimation and inference of the unknown parameters in this parametrization. By the curse of dimensionality, further modelling assumptions are typically necessary to obtain estimators with good practical performance. In this article, analogue to the current literature on marginal structural models introduced by Robins (e.g., Robins (2000a), Robins (2000b)), we propose new classes of inverse probability of censoring weighted (IPCW) estimators, double robust inverse probability of censoring weighted (DR-IPCW) estimators, and likelihood/regression-based estimators. (In our technical report van der Laan and Petersen (2004) we discussed plug-in estimators based on the identifiability result for the direct effect, which were not based on a model for the direct effect parameter.)

This article is organized as follows. In Section 2 we present our proposed models for direct effects, based on the statistical counterfactual framework as used by Robins and Greenland (1992) and Robins (2003), which assumes sequential randomization so that (total) causal effects are identifiable. In Section 3 we present the three methods for estimation. In Section 4 we present methods for statistical inference for the direct effect parameter. We end this article with a discussion.

For the sake of presentation, in the main part of this article we present our models and corresponding estimators of the direct effect of a point-treatment (that is, treatment is assigned at a single point in time and is not subject to subsequent random changes) followed by a single intermediate variable and outcome of interest. In the Appendix (Section 5) we generalize the statistical models and estimators to direct effects of time-dependent treatment regimens, not mediated by a specified time-dependent covariate process, based on general longitudinal data structures.

2 Direct Effect Models: The point treatment case.

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 co-variables 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 of interest. Let P_0 denote the sampling distribution of O.

2.1 The definition of direct effect in current literature.

In this subsection, we present the statistical framework and definitions of direct and indirect effects as presented in 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 structure observed on a randomly sampled 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 combination of possible treatment and possible intermediate variable. Instead, our observed data structure is only 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 counterfactuals Y_{az} (for the randomly sampled subject), and 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 missingness variable, and O is a specified function of X and A. Here $\mathcal{A}$ and $\mathcal{B}$ denote the support of A and (A, Z), respectively. 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 $P_0 = P_{F_{X0},g_0}$.

One defines the natural direct effect of changing treatment from 0 (e.g., representing a conventional treatment or no treatment) to a within strata of our sampling population defined by a baseline co-variate $V \subset W$ as

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

In order to identify controlled direct effects $E(Y_{az} - Y_{0z} \mid V)$ at a fixed level of the intermediate variable, we make the following randomization assumption:

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

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

$$A \perp (Z_a : a \in \mathcal{A}) \mid W. \tag{3}$$

Because of these randomization assumptions (3) and (2), we have the following relation between observed data probabilities and counterfactual probabilities:

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

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

$$E(Y_{az} - Y_{0z} \mid Z_0 = z, W) = E(Y_{az} - Y_{0z} \mid W) \text{ for all } (a, z) \in \mathcal{B}. \tag{4}$$

In the above model for the observed data distribution defined by (1), (2), (3), and (4), van der Laan and Petersen (2004) show that $E(Y_{aZ_0} - Y_{0Z_0} \mid V)$ equals

$$E_{W|V} \int_z \{E(Y \mid A = a, Z = z, W) - E(Y \mid A = 0, Z = z, W)\} P(Z = z \mid A = 0, W), \tag{5}$$

and thus that $E(Y_{aZ_0} - Y_{0Z_0} \mid V)$ is a (non-parametric) identifiable parameter.

2.2 A generalized class of direct effect parameters.

We argue that, even without the identifiability assumption (4), (5) is still an important direct effect parameter of interest, since, by the randomization assumptions only, (5) equals

$$DE(a,V) \equiv E\left(\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid V\right). \tag{6}$$

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 $Y_{az} - Y_{0z}$ w.r.t. to the conditional distribution of Z_0, given W. Therefore, if one is not comfortable with the identifiability assumption (4), then one can view the latter direct effect parameter $DE(a,V)$ as the parameter of interest, and our proposed estimators are estimators of $DE(a,V)$.

We actually wish to generalize this definition (6) of direct effect to also handle subject-specific weighted averages of the z-specific individual controlled direct effects w.r.t. to a user-supplied conditional distribution $Q_0(\cdot \mid W)$, given W. Therefore, we will define the parameter of interest as

$$DE(a,V) = DE(a,V \mid Q_0) \equiv E\left(\sum_z (Y_{az} - Y_{0z}) Q_0(z \mid W) \mid V\right), \tag{7}$$

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 (6).

2.3 Modelling the direct effect parameter.

Since $DE(a,V)$ is our parameter of interest, representing the answer to the scientific question of interest about our population, it is sensible practice, if $DE(a,V)$ is high dimensional (e.g., A or V is continuous, and/or high-dimensional), to model this function $DE(a,V)$. Consider a user-supplied parametrization/model $\beta \rightarrow m(a,V \mid \beta)$ for this direct effect parameter $DE(a,V)$ in terms of a Euclidean parameter β:

$$DE(a,V) = E(\sum_z (Y_{az} - Y_{0z}) Q_0(z \mid W) \mid V) = m(a,V \mid \beta_0). \tag{8}$$

This parametrization has to be chosen so that it satisfies $m(0,V \mid \beta) = 0$ for all V and β. The true β_0 represents now our 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 missing data structure

$$O = (W, A, Z, Y_{AZ}), \tag{9}$$

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 (2002)). We will denote the missing data model for P_0, defined by (9), (2), and (8) 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 $X = ((Z_a : a \in \mathcal{A}), (Y_{az} : (a, z) \in \mathcal{B}), W)$, where now only A plays the missingness variable, and in order to identify Q_0 one will also need the randomization assumption (3). So in this case the model for P_0 is defined by (1), (2), (3), and (8), and we will denote this 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 $\mathcal{M}(CAR)$.

Models for nuisance parameters: As we will see, the class of all estimating functions for β_0 in model $\mathcal{M}^*(CAR)$ is indexed by potentially high dimensional nuisance parameters so that the construction of asymptotically linear estimators requires specification of models for these nuisance parameters. That is, in order to deal with the curse of dimensionality in the model $\mathcal{M}^*(CAR)$, depending on the choice of class of estimators, we will also need to assume models for 1) $P_{A|W}$, $P_{Z|A,W}$, or equivalently, the missingness mechanism $P_{(A,Z)|X^*}$ (IPCW, DR-IPCW), and 2) $E(Y \mid A, Z, W)$ (Likelihood-based, DR-IPCW). In the case that $Q_0 = P_{Z_0|W}$, and is thus unknown, then the consistency for all three classes of estimators relies upon a consistent estimator of $P_{Z|A,W}$. In addition, the consistency of the IPCW-estimators relies upon a consistent estimator of $P_{A|W}, P_{Z|A,W}$, the consistency of the Likelihood-based estimators rely upon a consistent estimator of $E(Y \mid A, Z, W)$, while the consistency of the DR-estimators relies upon either a consistent estimator of $P_{A|W}, P_{Z|A,W}$ or a consistent estimator of $E(Y \mid A, Z, W)$ (but it uses both estimators).

Working model: Our estimators also require a specification of a *working* model $\{m_0(V \mid \eta) : \eta\}$ for

$$m_0(V) \equiv E(\sum_z Y_{0z} Q_0(z \mid W) \mid V),$$

and a corresponding estimator. However, the validity of this working model for m_0 (i.e., the consistency of the corresponding estimator of m_0) does only potentially affect the efficiency of our estimators of β_0, but it does *not* affect the consistency and asymptotic linearity of our estimators.

Given user-supplied models for these nuisance parameters, we can use the maximum likelihood estimator defined as the maximizers over these models of the relevant partial likelihoods given by

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

Other (e.g., estimating function-based) procedures for estimation of the nuisance parameters can be used as well, and, in particular, one could use cross-validation methodology to data-adaptively select the models for these parameters.

3 Estimation.

This section is organized as follows. In the next two subsections we present the IPCW estimating functions and corresponding estimators. Subsequently, we present the more general class of DR-IPCW estimating functions, and corresponding DR-IPCW estimators. We also present likelihood-based estimators. Finally, we discuss the properties of these three classes of estimators.

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: note, $g_0(A,Z \mid X^*) = g_0(A \mid W) g_0(Z \mid A, W)$. Recall that $Q_0(\cdot \mid W)$ is either user supplied and known, or equals the unknown $P_{Z_0|W} = P_{Z|A=0,W}$.

Consider the following class of inverse of probability of censoring weighted (IPCW) estimating functions for β_0 indexed by a user supplied functions $h(A,V) = (h_1(A,V), h_2(V), g^*(A \mid V))$ of A, V, and nuisance parameter g:

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

Here $g^*(\cdot \mid V)$ can be any user supplied conditional density of A, given V. The following lemma establishes that these estimating functions are indeed unbiased for β_0 at a correctly specified censoring mechanism g_0.

Lemma 1 *In addition to assuming model $\mathcal{M}^*(CAR)$ for P_0, we also assume the following experimental treatment assignment assumptions for the joint "treatment" (A,Z):*

$$\max_{(a,z)\in\mathcal{B}} \frac{h_1(a,V)}{g_0(a,z \mid W)} < \infty \ a.e.. \quad (10)$$

Then for any function $h = (h_1, h_2, g^*)$ of A, V

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

We also have that, if $h_2(V) = m_0(V) = E(\sum_z Y_{0z} Q_0(z \mid W) \mid V)$, then

$$E(D_{h,IPCW}(O \mid \beta_0, g_0, Q_0) \mid X^*) = \sum_a g^*(a \mid V)\{h_1(a,V) - E_{g^*}(h_1(A,V) \mid V)\}\{\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) - m(a,V \mid \beta_0)\}. \quad (11)$$

Remark regarding (11). In the full data model for $X^* = ((Y_{az} : (a,z) \in \mathcal{B}), W)$ defined by only the restriction (8), $E(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) \mid V) = m(a, V \mid \beta_0)$ for some β_0, the orthogonal complement of the nuisance tangent space at P_0 for β_0 (for known Q_0) is given by:

$$T^{F,\perp}_{nuis}(F_{X^*0}) = \left\{ \sum_a h(a,V) \left(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) - m(a,V \mid \beta_0) \right) : h \right\}.$$

This follows from the fact that this full data model (8) is simply a repeated measures regression model for the outcome vector ($H_a = \sum_z Q_0(z \mid$

$W)(Y_{az} - Y_{0z}) : a)$ on V, and application of Theorem 2.2 in van der Laan and Robins (2002). Therefore, the latter property in Lemma 1 shows that the conditional expectations of the IPCW-estimating functions, given the full data X^*, contain the orthogonal complement of the nuisance tangent space in the full data model. The latter property formally proves that the class of estimating functions $\{D_{h,IPCW}; h\}$ are indeed IPCW estimating functions as defined in van der Laan and Robins (2002). This property teaches us that our augmented class of DR-IPCW estimating functions as presented in the next subsection provide a representation of the orthogonal complement of the nuisance tangent space at P_0 of β_0 in the observed data model $\mathcal{M}^*(CAR)$: see Theorem 1.3, page 64 in van der Laan and Robins (2002).

Proof of Lemma 1: For notational convenience, we suppress the "IPCW" labelling. Firstly, we condition on $X^* = (Y_{az} : a, z), W)$, which corresponds with integrating over A, Z w.r.t. $g_0(A, Z \mid W)$. This yields

$$\begin{aligned}
&E(D_h(O \mid \beta_0, g_0, Q_0) \mid X^*) = \\
&\textstyle\sum_{a,z} g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V))Q_0(z \mid W)(Y_{az} - m(a, V \mid \beta_0) - h_2(V)) \\
&= \textstyle\sum_a g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V)) \left(\sum_z Q_0(z \mid W)Y_{az} - m(a, V \mid \beta_0) - h_2(V)\right) \\
&= \textstyle\sum_a g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V)) \times \\
&\left(\textstyle\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) - m(a, V \mid \beta_0) + m_0(V) - h_2(V)\right),
\end{aligned}$$

where we recall that $m_0(V) = E(\sum_z Q_0(z \mid W)Y_{0z} \mid V)$. At the first equality, we relied on the ETA (10) 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,

$$\sum_a g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V))\{m_0(V) - h_2(V)\} = 0,$$

which completes the proof of Lemma 1. □

3.2 Inverse probability of censoring weighted estimators.

Estimation of the index: A convenient choice for the function h indexing the IPCW-estimating functions $D_{h,IPCW}$ is given by

$$\begin{aligned}
h_1^*(A, V) &\equiv \frac{d}{d\beta_0} m(A, V \mid \beta_0) \\
h_2^*(V) &= m_0(V) = E_0(\sum_z Y_{0z} Q_0(z \mid W) \mid V) \\
g^*(A \mid V) &= g_0(A \mid V).
\end{aligned}$$

Let h_n be an estimator of this function $h*$ 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

$$m_0(V) = E_0(\sum_z Q_0(z \mid W) E_0(Y \mid A = 0, Z = z, W) \mid V),$$

we have that h_{2n} can be obtained as a regression estimator 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 the working model $m(V \mid \eta)$ for m_0. In case $m(A, V \mid \beta)$ 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 guessed β.

The corresponding IPCW-estimator: In addition, let g_n be an estimator of the missingness/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:

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

Since this estimator is a special case of the double robust IPCW estimator, we discuss implementation of this estimator in the next subsection.

Weighted least squares IPCW-estimators under a correctly specified model for m_0. At cost of robustness w.r.t. misspecification of $m_0(V)$, we can proposes a class of IPCW-estimators which can be represented as a weighted least squares estimator. These estimators are based on the observation that $m(\cdot \mid \beta)$ and $m(V \mid \eta)$ imply the model $E(\sum_z Q_0(z \mid W)Y_{az} \mid V) = m_1(a, V \mid \theta_0) \equiv m(a, V \mid \beta_0) + m(V \mid \eta_0)$, where now $\theta_0 = (\beta_0, \eta_0)$ represents the parameter of interest. This full data repeated measures regression model suggests as class of IPCW-estimating functions for θ_0:

$$\left\{ \frac{g^*(A \mid V)}{g(A, Z \mid X^*)} h(A, V) Q_0(Z \mid W)(Y - m_1(A, V \mid \theta)) : h \right\}.$$

It is straightforward to verify that these estimating functions are indeed unbiased for θ_0 at a correctly specified g_0. The IPCW-estimator indexed by the choice $h(A, V) = d/d\theta m_1(A, V \mid \theta)$ minimizes a weighted sum of squared residuals of the regression of Y on A, V based on the model $m(A, V \mid \theta)$. Specifically, this estimator is defined as

$$\theta_{n,IPCW} \equiv \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - m_1(A_i, V_i \mid \theta))^2 \text{weight}_i,$$

where

$$\text{weight}_i \equiv \frac{g_n^*(A_i \mid V_i) Q_{0n}(Z_i \mid W_i)}{g_n(A_i \mid W_i) Q_{Zn}(Z_i \mid A_i, W_i)}.$$

As a consequence, this IPCW-estimator $\theta_{n,IPCW}$ can be implemented with standard regression software. Since this estimator is not protected against misspecification of $m_0(V)$ (in other words, the working model $m(V \mid \eta)$ needs to be correctly specified), in the case that the parameter of interest is only β_0, and not both (β_0, η_0), then we only recommend to use this weighted least squares regression estimator as a starting value for solving the wished IPCW estimating equation.

Remark. In model $\mathcal{M}^*(CAR)$, assuming Q_0 is known, it follows from the general Theorem 2.3 in van der Laan and Robins (2002) that the asymptotic efficiency of $\beta_{n,IPCW}$ improves if we estimate the missingness mechanism $g_0(A, Z \mid X^*)$ more nonparametrically.

In the model $\mathcal{M}(CAR)$ in which $Q_0 = P_{Z_0|W}$, X is the full data structure, and A is the missingness variable, this same Theorem 2.3 teaches us that $\beta_{n,IPTW}$'s efficiency improves if we estimate the missingness/treatment mechanism $g_0(A \mid W)$ more nonparametrically. For example, if treatment is randomized so that $g_0(A \mid W) = g_0(A)$ is known, then the IPCW estimator using an estimator of $g_0(A)$ based on a logistic regression model including co-variables extracted from W will be significantly more efficient than the estimator using the known g_0. Since in this case the parameter of interest β_0 is not variation independent of $P_{Z|A,W}$, it is unclear how the size of the model for $P_{Z|A,W}$ affects the efficiency of the resulting estimator $\beta_{n,IPCW}$.

3.3 Double robust IPCW-estimating functions.

Given a class of IPCW-estimating functions, van der Laan and Robins (2002) (Theorem 1.3) show that the class of all (i.e., relevant) estimating functions in model $\mathcal{M}^*(CAR)$ is obtained 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 inverse probability of missingness weighted 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 $\langle f_1, f_2 \rangle_{P_0} \equiv E_0 f_1(O) f_2(O)$ being the covariance operator. Thus these estimating function are derived as $D_{h,DR} = D_{h,IPTW}(O) - \{E(D_{h,IPTW}(O) \mid A, Z, W) - E(D_{h,IPTW}(O) \mid W)\}$.

This class of estimating functions represents all estimating functions in the sense that $\{D_{h,DR} : h\} \subset L_0^2(P_0)$, evaluated at the true parameter values, contains the orthogonal complement of the nuisance tangent space at P_0 of β_0 in model $\mathcal{M}^*(CAR)$. This implies that the corresponding class of estimating equations generates all regular asymptotically linear estimators of β_0 up till a second order term. This follows from Theorem 1.3 in van der Laan and Robins (2002) and the established property (11) of the IPCW-estimating functions. This means that there also exists a h_{opt} so that $D_{h_{opt,DR}}$ is an optimal estimating function resulting in an efficient estimator if the nuisance parameters are correctly specified. Estimation of h_{opt} would result in locally efficient estimators, but that is beyond the scope of this article.

We have that

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

Thus,

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

If we let $Q_Y(A, Z, W)$ represent a parameter value for $Q_{Y0}(A, Z, W) = E_0(Y \mid A, Z, W)$, then we have

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

As predicted by the general estimating function theory (Section 1.6, van der Laan and Robins (2002)), these estimating functions are double robust w.r.t. the pair of nuisance parameters (g_0, Q_{Y0}).

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

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

If (10) holds at g, then for any index h

$$ED_{h,DR}(O \mid \beta_0, g, Q_Y, Q_0) = 0 \text{ if either } g = g_0 \text{ or } Q_Y = Q_{Y0}. \tag{12}$$

This can be straightforwardly explicitly verified. That is, this estimating function for β_0, which is indexed by two nuisance parameters, is unbiased if the ETA (10) holds at the possibly misspecified missingness-mechanism g, and one the two nuisance parameters $g_0 = p_{A,Z|W}$, $Q_{Y0} = E_0(Y \mid A, Z, W)$ is correctly specified as well. In practice, the requirement that the ETA (10) holds at g translates into using an estimator of g_0 which satisfies the ETA (10) (which can always be arranged).

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}.

Implementation: As communicated to us by Dan Rubin in our Department through personal communication, if $m(\cdot \mid \beta)$ is linear in β, then this estimating equation is just a linear system of equations in β, and can thus be solved trivially in closed form. For general parameterizations this estimator can be computed with the Newton-Raphson algorithm, and 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 $\beta_{n,IPTW}$ as initial estimator. For more details, we refer to van der Laan and Robins (2002) (pages 118-119).

3.5 A likelihood-based estimator.

Consider the identifiability result (5) for $DE(a,W) = E(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) \mid W)$ given by:

$$DE(a,W) = \int_z \{E(Y \mid A = a, Z = z, W) - E(Y \mid A = 0, Z = z, W)\} Q_0(z \mid W). \tag{13}$$

The fact that $E_0(DE(a,W) \mid V) = m(a,V \mid \beta_0)$ suggests to plug in an estimator Q_{Yn} of $Q_{Y0}(A,Z,W) = E(Y \mid A,Z,W)$ to obtain a fitted $\hat{DE}(a,W)$, and subsequently regress the vector $(\hat{DE}(a,W_i) : a)$, on V according to the repeated measured regression 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), and are therefore not repeated here.

3.6 Discussion of the three types of estimators.

The consistency of the IPCW-estimator of β_0 relies on the consistency of g_n as an estimator of the missingness mechanism $g_0(\cdot \mid X^*)$, and on the experimental "treatment" assumption (10) on g_0 jointly for (A,Z). 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 (10) 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_{0Y}, but, if (10) fails to hold for g_0, then it fully relies on consistent estimation of Q_{0Y}.

The double robust estimator has the attractive property as being the most nonparametric estimator, which will be consistent if either the IPTW-estimator or the likelihood-based estimator is consistent. In the case that one expects a serious violation of the joint ETA (10), then the likelihood-based estimator might be the preferred estimator, since the double robust estimator now fully relies on the consistent estimation of $E(Y \mid A,Z,W)$ as well. That is, in the latter case, the DR estimator is not more robust than the likelihood-based estimator. The bias caused by the violation of the joint ETA assumption (10) can be established by computing the sampling distribution of the IPCW-estimator under a maximum likelihood estimator of the data generating distribution (i.e., one implements a parametric bootstrap), and comparing the mean of the sampling distribution with the β_0 as calculated from the fitted likelihood. Off course, the latter parametric bootstrap can also be used to estimate the sampling distribution of the DR-IPTW estimator (and thereby its mean squared error and bias), and it can be used as a model based inference tool to construct confidence intervals.

4 Statistical Inference.

Consider the case that Q_0 is a known weight function. Under the above stated assumptions regarding correct estimation of the nuisance parameters g_0, Q_{Y0}, and regularity conditions guaranteeing that 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 also be shown that the IPTW and DR-IPTW 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 (2002). Under these regularity conditions, one can also establish the asymptotic validity of the bootstrap for obtaining confidence regions for β_0. These confidence regions can also be used for testing purposes.

We will discuss here the conclusions of this Theorem 2.4 in van der Laan and Robins (2002) which relies on assuming a correctly specified model for g_0. Since the IPCW-estimating functions correspond with setting $Q_Y = 0$ in the double robust estimating functions, it suffices to present the statements for the double robust estimator of β_0, which we denote with β_n here. Firstly, in the unrealistic case that $g_n = g_0$, then under regularity conditions specified in Theorems 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 missingness 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 missingness mechanism g_0), then Theorem 2.4 in van der Laan and Robins (2002) 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 (2002)).

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 missingness mechanism scores (i.e., $T_{CAR}(P_0)$), so that $IC = IC_0$ (i.e., the projection now equals zero). In general, 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 we already need in our construction of β_n, and it avoids the computer intensive re-sampling approach. It is our experience that this approach is not very conservative at all for the DR-IPCW estimators, assuming one does a reasonable job in estimating Q_{Y0}, but that it can be very conservative for the IPCW estimator, which corresponds with extreme misspecification of Q_{Y0}. That is, in the case one uses the IPCW-estimator it is really worthwhile to calculate the projection component onto the tangent space of the missingness mechanism model of the true influence curve (except if one knows and uses (unwisely) the true g_0).

We suggest that, even in the double robust model (Theorem 2.5 in van der Laan and Robins (2002)) 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 provides typically more accurate estimates of the *finite* sample distribution than the influence curve (Wald-type) approach described above.

In the case that the weight function 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 explicitly calculated. In case one wishes to avoid such calculations, we suggest to to simply use the bootstrap.

5 Discussion.

The aim of this article is to give direct effect estimation in the causal inference literature the same place as estimation of total causal effects through (e.g.) marginal structural models. In fact, our approach has been to model the observed data distribution with the standard causal model treating A, Z as a joint treatment, only assuming sequential randomization, but we defined our parameter of interest as a new kind of causal parameter in this model. By adding the conditional independence assumption of van der Laan and Petersen (2004) and selecting as weight function $Q_0 = P_{Z_0|W}$ this parameter happens to reduce to the conventional definition of direct effect, but it remains an interesting direct effect parameter if this assumption fails to hold. Our proposed causal parameter is also meaningful in the case that Z represents the role of another treatment component (say) A_2. In this case our direct effect parameter represents a population mean of a subject specific average of a_2-specific treatment effects $Y_{a,a_2} - Y_{0,a_2}$, fixing the other component of treatment, over a_2 w.r.t. to a conditional distribution of a_2 indexed by observed baseline characteristics of the subject.

Acknowledgement.

We thank James Robins for helpful discussions. This research was funded by NIH grant R01-AI550850 and R01-GM071397. Maya Petersen is supported by a Pre-doctoral Fellowship from the Howard Hughes Medical Institute.

Appendix: Generalization to time-dependent treatments.

Consider the longitudinal data structure

$$O = (W = L(0), A(0), Z(0), \ldots, L(K), A(K), Z(K), Y = L(K+1)),$$

and suppose that we observe n i.i.d. copies of O. We will use the notation $A = (A(0), \ldots, A(K))$, $Z = (Z(0), \ldots, Z(K))$, and for a time-dependent process (say) X, we denote its history with $\bar{X}(t) = (X(0), \ldots, X(t))$.

We assume the existence of a random variable $X \equiv ((Z_a : a \in \mathcal{A}), (L_{az} : (a,z) \in \mathcal{B}))$ of treatment specific counterfactuals Z_a and counterfactuals L_{az} (for the randomly sampled subject), and that

$$O = (W = L(0), A(0), Z_A(0), \ldots, L_{A,Z}(K), A(K), Z_A(K), Y_{A,Z} = L_{A,Z}(K+1)) \tag{14}$$

is a missing data structure on the full data structure X. That is, X is the full data structure of interest, A is the missingness variable, and O is a specified function of X and A. The temporal ordering assumption states that $L_{A,Z}(j) = L_{\bar{A}(j-1),\bar{Z}(j-1)}(j)$. Here $\mathcal{A}$ and $\mathcal{B}$ denote the support of A and (A,Z), respectively.

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 $P_0 = P_{F_{X0},g_0}$.

One defines the natural direct effect of changing treatment from 0 (e.g., representing a conventional treatment or no treatment) to a within strata of our population defined by a baseline co-variate $V \subset W$ as

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

In order to identify controlled direct effects $E(Y_{az} - Y_{0z} \mid V)$ at a fixed level of the intermediate variable, we assume sequential randomization of the joint $(A(j), Z(j))$:

$$A(j), Z(j) \perp X^* \equiv (L_{az} : (a,z)) \mid \bar{A}(j-1), \bar{Z}(j-1), \bar{L}(j). \tag{15}$$

In order to identify a conditional distribution of Z_a, given W, we assume the sequential randomization assumption for treatment:

$$A(j) \perp X \mid \bar{A}(j-1), \bar{Z}(j-1), \bar{L}(j). \tag{16}$$

Specifically, because of these randomization assumptions (16) and (15), we have the following G-computation identifiability results for the marginal distributions of L_{az} and $(Z_a, L_a = L_{aZ_a})$:

$$
\begin{aligned}
P_{L_{az}}(L) &= \prod_j P(L(j) \mid \bar{L}(j-1), \bar{A}(j-1) = \bar{a}(j-1), \bar{Z}(j-1) = \bar{z}(j-1)) \\
P_{Z_a, L_a}(Z, L) &= \prod_j P(Z(j), L(j) \mid \bar{Z}(j-1), \bar{L}(j-1), \bar{A}(j-1) = \bar{a}(j-1)).
\end{aligned}
$$

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

$$
E(Y_{az} - Y_{0z} \mid Z_0 = z, W) = E(Y_{az} - Y_{0z} \mid W) \text{ for all } (a, z) \in \mathcal{B}. \tag{17}
$$

In the above model for the observed data distribution defined by (14), (15), (16), and (17), van der Laan and Petersen (2004) show that $E(Y_{aZ_0} - Y_{0Z_0} \mid V)$ equals

$$
E_{W|V} \int_z \{E(Y_{az} \mid W) - E(Y_{0z} \mid W)\} P(Z_0 = z \mid W), \tag{18}
$$

and thus, by the G-computation formulas above, that $E(Y_{aZ_0} - Y_{0Z_0} \mid V)$ is a (non-parametric) identifiable parameter.

A generalized class of direct effect parameters.

We argue that, even without the identifiability assumption (17), (18) is still an important direct effect parameter of interest, since, by the sequential randomization assumptions only, we have that (18) equals

$$
DE(a, V) \equiv E\left(\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid V \right). \tag{19}
$$

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 $Y_{az} - Y_{0z}$ w.r.t. to the conditional distribution of Z_0, given W. Therefore, if one is not comfortable with the identifiability assumption (17), then one can view the latter direct effect parameter $DE(a, V)$ as the parameter of interest, and our proposed estimators are estimators of $DE(a, V)$.

We actually wish to generalize this definition (19) of direct effect to also handle subject-specific weighted averages of the z-specific individual controlled direct effects w.r.t. to a user-supplied conditional distribution $Q_0(\cdot \mid W)$, given W. Therefore, we will define the parameter of interest as

$$DE(a,V) = DE(a,V \mid Q_0) \equiv E\left(\sum_z (Y_{az} - Y_{0z}) Q_0(z \mid W) \mid V\right), \quad (20)$$

where Q_0 could be known, or it could be the unknown $P(Z_0 = z \mid W)$ in which case this definition reduces to (6).

Modelling the direct effect parameter.

Since $DE(a,V)$ is our parameter of interest, representing the answer to the scientific question of interest about our population, it is sensible practice, if $DE(a,V)$ is high-dimensional (e.g., A or V is continuous, and/or high-dimensional), to model this function $DE(a,V)$. Consider a user-supplied parametrization/model $\beta \rightarrow m(a,V \mid \beta)$ for this direct effect parameter $DE(a,V)$ in terms of a Euclidean parameter β:

$$DE(a,V) = E(\sum_z (Y_{az} - Y_{0z}) Q_0(z \mid W) \mid V) = m(a,V \mid \beta_0). \quad (21)$$

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

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 $X* \equiv (L_{az} : (a,z) \in \mathcal{B})$, and (A,Z) can now be viewed as the joint missingness variable defining the observed missing data structure

$$O = (A, Z, L_{AZ}), \quad (22)$$

where we assume that (A,Z) is sequentially randomized as defined by (15), or equivalently, that this joint missingness mechanism satisfies coarsening at random (van der Laan and Robins (2002)). We will denote the missing data model for P_0, defined by (22), (15), and (8) 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 $X = ((Z_a : a \in \mathcal{A}), (L_{az} : (a,z) \in \mathcal{B}))$, where now A plays the missingness variable, and in order to identify Q_0 one will also need the randomization assumption (16). So in this case the model for P_0 is defined by (14), (15), (16), and (8), and we will denote this 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 $\mathcal{M}(CAR)$.

Models for nuisance parameters: As we will see the class of all estimating functions for β_0 in model $\mathcal{M}^*(CAR)$ is indexed by potentially high dimensional nuisance parameters so that the construction of asymptotically linear estimators requires specification of models for these nuisance parameters. That is, in order to deal with the curse of dimensionality in model $\mathcal{M}^*(CAR)$, depending on the choice of class of estimators, we will also need to assume models for 1) $g_0(A, Z \mid X^*)$ (IPCW, DR-IPCW), and 2) $Q_{L0} = \prod_j P(L(j) \mid \bar{L}(j-1), \bar{A}(j-1), \bar{Z}(j-1))$ (Likelihood-based, DR-IPCW). Obviously, in the case that $Q_0 = p_{Z_0|W}$, and is thus unknown, then the consistency for all three classes of estimators relies upon a consistent estimator of Q_0. In addition, the consistency of the IPCW estimators relies upon a consistent estimator of $g_0(A, Z \mid X^*)$, the consistency of the Likelihood-based estimators rely upon a consistent estimator of Q_{L0}, while the consistency of the DR estimators relies upon either a consistent estimator of g_0 or Q_{L0} (but it uses both estimators).

Working model: Our estimators also require a specification of a *working* model $\{m_0(V \mid \eta) : \eta\}$ for $m_0(V) \equiv E(\sum_z Y_{0z} Q_0(z \mid W) \mid V)$, and a corresponding estimator. We have

$$m_0(V) = E\left(\sum_z Q_0(z \mid W) E(Y_{0z} \mid W) \mid V\right).$$

The validity of this working model for m_0 does only potentially affect the efficiency of our proposed IPTW and DR-IPTW estimators of β_0, but it does *not* affect the consistency and asymptotic linearity of our estimators.

The missingness mechanism g could be modelled as $g_1 * g_2$, by noting that

$$g(A, Z \mid X) = \prod_j g_1(A(j) \mid \bar{A}(j-1), \bar{Z}(j-1), \bar{L}(j)) \prod_j g_2(Z(j) \mid \bar{A}(j), \bar{Z}(j-1), \bar{L}(j)).$$

Given user-supplied models for these nuisance parameters ($g = (g_1, g_2), Q_L$), we can use the maximum likelihood estimator defined as the maximizers over these models of the relevant partial likelihoods given by

$$L(g) = \prod_{i=1}^{n}\prod_{j} g(A_i(j), Z_i(j) \mid \bar{A}_i(j-1), \bar{Z}_i(j-1), \bar{L}_i(j))$$

$$L(Q_L) = \prod_{i=1}^{n}\prod_{j} Q_L(L_i(j) \mid \bar{L}_i(j-1), \bar{A}_i(j-1), \bar{Z}_i(j-1)).$$

Other (e.g., estimating function-based) procedures for estimation of the nuisance parameters can be used as well, and, in particular, one could use cross-validation methodology to data-adaptively select the models for these parameters.

Estimation.

This section is organized as follows. In the next two subsections we present the IPCW estimating functions and corresponding estimators. Subsequently, we present the more general class of DR-IPCW estimating functions, and DR-IPCW estimators. We also present likelihood-based estimators.

Inverse Probability of Treatment 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. Recall that $Q_0(\cdot \mid W)$ is either user-supplied and known, or equals the unknown $P_{Z_0|W} = P_{Z|A=0,W}$.

Consider the following class of inverse of probability of treatment weighted (IPCW) estimating functions for β_0 indexed by a user-supplied functions $h(A, V) = (h_1(A, V), h_2(V), g^*(A \mid V))$ of A, V, and nuisance parameter g:

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

Here $g^*(\cdot \mid V)$ can be any user-supplied conditional density of $A = (A(0), \ldots, A(K))$, given V. The following lemma establishes that these estimating functions are indeed unbiased.

Lemma 2 *In addition to assuming model $\mathcal{M}^*(CAR)$ for P_0, we also assume the following experimental treatment assignment assumptions for the joint* *"treatment" (A, Z):*

$$\max_{(a,z)\in\mathcal{B}} \frac{h_1(a,V)}{g_0(a,z \mid X^*)} < \infty \ a.e.. \tag{23}$$

Then for any function h

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

We also have that, if $h_2(V) = m_0(V) = E(\sum_z Y_{0z} Q_0(z \mid W) \mid V)$, then

$$\begin{aligned} E(D_{h,IPCW}(O \mid \beta_0, g_0, Q_0) \mid X^*) &= \qquad (24) \\ \sum_a g^*(a \mid V)(h_1(a,V) - E_{g^*}(h_1(A,V) \mid V)) &\left(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) - m(a, V \mid \beta_0) \right). \end{aligned}$$

Remark regarding (24). In the full data model for $X^* = ((L_{az} : (a,z) \in \mathcal{B}), W)$ defined by the only restriction (8), $E(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) \mid V) = m(a, V \mid \beta_0)$ for some β_0, the orthogonal complement of the nuisance tangent space at P_0 for β_0 (for known Q_0) is given by:

$$T^{F,\perp}_{nuis}(F_{X^*0}) = \left\{ \sum_a h(a,V) \left(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) - m(a, V \mid \beta_0) \right) : h \right\}.$$

This follows from the fact that this full data model (8) is simply a repeated measures regression model for the outcome vector $(H_a = \sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) : a)$ on V, and application of Theorem 2.2 in van der Laan and Robins (2002). Therefore, the latter property in Lemma 2 shows that the conditional expectations of the IPCW-estimating functions at the true parameter values, given X^*, contain the orthogonal complement of the nuisance tangent space in the full data model. The latter property formally proves that the class of estimating functions $\{D_{h,IPCW}; h\}$ are indeed IPCW estimating functions as defined in van der Laan and Robins (2002). This property teaches us that our augmented class of DR-IPCW estimating functions as presented in the next subsection provide a representation of the orthogonal complement of the nuisance tangent space at P_0 of β_0 in the observed data model $\mathcal{M}^*(CAR)$: by Theorem 1.3, page 64 in van der Laan and Robins (2002).

Proof of Lemma 2: For notational convenience, we suppress the "IPCW" labelling. Firstly, we condition on $X^* = (Y_{az} : a, z), W)$, which corresponds with integrating over A, Z w.r.t. $g_0(A, Z \mid X^*)$. This yields

$$
\begin{aligned}
&E(D_h(O \mid \beta_0, g_0, Q_0) \mid X^*) = \\
&\sum_{a,z} g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V))Q_0(z \mid W)\times \\
&(Y_{az} - m(a, V \mid \beta_0) - h_2(V)) \\
&= \sum_a g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V))\times \\
&(\sum_z Q_0(z \mid W)Y_{az} - m(a, V \mid \beta_0) - h_2(V)) \\
&= \sum_a g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V))\times \\
&(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) - m(a, V \mid \beta_0) + m_0(V) - h_2(V)),
\end{aligned}
$$

where we recall that $m_0(V) = E(\sum_z Q_0(z \mid W)Y_{0z} \mid V)$. At the first equality, we relied on the ETA (23) 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,

$$\sum_a g^*(a \mid V)(h_1(a, V) - E_{g^*}(h_1(A, V) \mid V))\{m_0(V) - h_2(V)\} = 0,$$

which completes the proof of Lemma 2. □

Inverse probability of censoring weighted estimators.

Estimation of the index: A convenient choice for the function $h = (h_1, h_2, g^*)$ indexing the IPCW-estimating functions $D_{h,IPCW}$ is given by

$$
\begin{aligned}
h*_1(A, V) &\equiv \frac{d}{d\beta_0} m(A, V \mid \beta_0) \\
h_2^*(V) &= m_0(V) = E_0(\sum_z Y_{0z} Q_0(z \mid W) \mid V) \\
g^*(A \mid V) &= p_0(A \mid V).
\end{aligned}
$$

Let h_n be an estimator of this function $h*$ based on substitution of an estimator g_n^* of the conditional distribution of A, given V, and a regression estimator $h_{2n}(V)$ of m_0. If $Q_0(z \mid W) = p_{Z_0 \mid W}$, then, the estimate h_{2n} can be chosen to be an estimator of $E(Y_0 \mid V)$ (and thus of $m_0(V)$ under the conditional independence assumption). In general, we have

$$m_0(V) = E_0(\sum_z Q_0(z \mid W)E_0(Y_{0z} \mid W) \mid V),$$

so that h_{2n} can be obtained as a regression estimator obtained by regressing $\sum_z Q_0(z \mid W_i)\hat{E}(Y_{0z} \mid W_i)$ on V_i according to the working model $m(V \mid \eta)$ for

m_0. Estimation of $E(Y_{az} \mid W)$ can be carried out with marginal structural model methodology (e.g., van der Laan and Robins (2002), Chapter 6). In case $m(A, V \mid \beta)$ 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 a h_1^* at an initial guess of β_0.

The corresponding IPCW-estimator: In addition, let g_n be an estimator of the missingness mechanism $g_0(A, Z \mid X^*)$. 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:

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

Weighted least squares IPCW-estimators under a correctly specified model for m_0. With some modifications at cost of robustness w.r.t. misspecification of $m_0(V)$, we can also propose IPCW-estimators which can be represented as weighted least squares estimators. Firstly, we observe that $m(\cdot \mid \beta)$ and $m(V \mid \eta)$ implies the following model $E(\sum_z Q_0(z \mid W) Y_{az} \mid V) = m_1(a, V \mid \theta_0) \equiv m(a, V \mid \beta_0) + m(V \mid \eta_0)$, where now $\theta_0 = (\beta_0, \eta_0)$ represents the parameter of interest. This full data repeated measures regression model suggests as class of IPCW-estimating functions for θ_0:

$$\left\{ \frac{g^*(A \mid V)}{g(A, Z \mid X^*)} h(A, V) Q_0(Z \mid W)(Y - m_1(A, V \mid \theta)) : h \right\}.$$

It is straightforward to verify that these estimating functions are indeed unbiased for θ_0 at a correctly specified g_0. The IPCW-estimator indexed by the choice $h(A, V) = d/d\theta m_1(A, V \mid \theta)$ minimizes a weighted sum of squared residuals of the regression of Y on A, V based on the model $m(A, V \mid \theta)$. Specifically, this estimator is defined as

$$\theta_{n,IPCW} \equiv \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - m_1(A_i, V_i \mid \theta))^2 \text{weight}_i,$$

where

$$\text{weight}_i \equiv \frac{g_n^*(A_i \mid V_i) Q_{0n}(Z_i \mid W_i)}{g_n(A_i, Z_i | X^*)}.$$

As a consequence, this IPCW-estimator $\theta_{n,IPCW}$ can be implemented with standard regression software. Since this estimator is not protected against misspecification of $m_0(V)$ (in other words, the working model $m(V \mid \eta)$ needs to be correctly specified), in the case that the parameter of interest is only β_0, and not both (β_0, η_0), then we recommend to only use this weighted least squares regression estimator as a starting value for solving the wished IPCW estimating equation.

Remark. In model $\mathcal{M}^*(CAR)$, assuming Q_0 is known, it follows from the general Theorem 2.3 in van der Laan and Robins (2002) that the asymptotic efficiency of $\beta_{n,IPCW}$ improves if we estimate the missingness mechanism $g_0(A, Z \mid X^*)$ more non-parametric.

In the model $\mathcal{M}(CAR)$ in which $Q_0 = P_{Z_0|W}$, X is the full data structure, and A is the missingness variable, this same Theorem 2.3 teaches us that $\beta_{n,IPTW}$'s efficiency improves if we estimate the missingness/treatment mechanism $g_1(A \mid X) = \prod_j g_1(A(j) \mid \bar{A}(j-1), X)$ under the SRA (16) more nonparametric. For example, if treatment is randomized so that $g_{10}(A \mid X) = g_{10}(A)$ is known, then the IPCW estimator using an estimator of $g_{10}(A)$ based on a logistic regression model for $g_{10}(A(j) \mid \bar{A}(j-1), \bar{Z}(j-1), \bar{L}(j))$, including co-variables extracted from the observed past, will be significantly more efficient than the estimator using the known treatment mechanism g_0. Since the dependence of the parameter of interest β_0 on $P_{Z_0|W}$ implies that it is not variation independent of the intermediate-variable mechanism g_{20}, it is unclear how the size of the model for g_{20} affects the efficiency of the resulting estimator $\beta_{n,IPCW}$.

Double robust IPCW-estimating functions.

Given a class of IPCW-estimating functions, van der Laan and Robins (2002) show that the class of all (i.e., relevant) estimating functions in model $\mathcal{M}^*(CAR)$ is obtained by subtracting from the IPCW estimating functions its projection on all possible nuisance scores corresponding with one dimensional fluctuations of the true missingness mechanism $g_0(A, Z \mid X^*) = g_{10}(O)g_{20}(O)$ only assuming CAR (i.e., (15). That is, one subtracts from the inverse probability of missingness weighted estimating function its projection on the sub-Hilbert space of functions of O with conditional mean zero, given X^*, 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 $\langle f_1, f_2 \rangle_{P_0} \equiv E_0 f_1(O) f_2(O)$ being the covariance operator.

The closed form projection is presented in van der Laan and Robins (2002) (e.g. Theorem 1.2), and is given by:

$$D_{h,SRA}(O \mid g, Q_L, Q_0) \equiv \sum_{j=0}^{K} E_{g,Q_L}(D_{h,IPTW}(O \mid \beta(Q_L), g, Q_0) \mid \bar{A}(j), \bar{Z}(j), \bar{L}(j)) \\ -E_{g,Q_L}(D_{h,IPTW}(O \mid \beta(Q_L), g, Q_0) \mid \bar{A}(j-1), \bar{Z}(j-1), \bar{L}(j)),$$

where the latter j-specific term can also be represented as the conditional expectation of the first term over $A(j), Z(j)$ w.r.t. $g_0(A(j), Z(j) \mid \bar{A}(j-1), \bar{Z}(j-1), X^*)$. Here Q_L represents a parameter value for Q_{L0}, and $\beta(Q_L)$ is the parameter value $\beta(P)$ under a distribution P of O with density $Q_L(O) * g_0(O \mid X^*)$.

Thus, the double robust estimating functions can be represented as

$$D_{h,DR}(O \mid \beta, g, Q_L, Q_0) = D_{h,IPTW}(O \mid \beta, g, Q_0) - D_{h,SRA}(O \mid g, Q_L, Q_0).$$

This class of estimating functions represents all estimating functions in the sense that $\{D_{h,DR}(\cdot \mid \beta_0, g_0, Q_{L0}, Q_0) : h\} \subset L_0^2(P_0)$ contains the orthogonal complement of the nuisance tangent space at P_0 of β_0 in model $\mathcal{M}^*(CAR)$. This follows from Theorem 1.3 in van der Laan and Robins (2002) and the established property (24) of the IPCW-estimating functions. As a consequence, this class of double robust estimating functions includes the optimal estimating function $D_{h_{opt},DR}$ which equals the efficient influence curve at its true parameter values, and, as a consequence, estimation of h_{opt} according to a guessed submodel, and using the corresponding double robust IPCW estimating equation resuls in a locally efficient double robust IPCW estimator of β_0. However, this is beyond the scope of this article.

As predicted by the general estimating function theory (Section 1.6, van der Laan and Robins (2002)), this estimating function is double robust w.r.t. the pair of nuisance parameters (g_0, Q_{L0}).

Result 2 *Consider the class of double robust IPCW estimating functions:*

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

If (23) holds at g, then for any index h

$$ED_{h,DR}(O \mid \beta_0, g, Q_L, Q_0) = 0 \text{ if either } g = g_0 \text{ or } Q_L = Q_{L0}. \tag{25}$$

This is proved in van der Laan and Robins (2002) (Section 1.6). That is, this estimating function for β_0, which is indexed by two nuisance parameters, is unbiased if the ETA (23) holds at the possibly misspecified missingness-mechanism g, and one the two nuisance parameters $g_0 = p_{A,Z|X^*}$, Q_{L0} is correctly specified as well. In practice, the requirement that the ETA (23) holds at g translates into using an estimator of g_0 which satisfies the ETA (23) (which can always be arranged).

Double robust IPCW-estimators.

Let h_n, g_n, Q_{Ln} be estimators of h^*, g_0, Q_{L0}. The corresponding DR-IPCW-estimator of β_0 is now defined as the solution $\beta_{n,DR}$ of the estimating equation in β:

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

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

Implementation: If $m(\cdot \mid \beta)$ is linear in β, then this estimating equation is just a linear system of equations in β, and can thus be solved in closed form. For non-linear parameterizations $\beta \rightarrow m(\cdot \mid \beta)$, this estimator can be computed with the Newton-Raphson algorithm, and 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 $\beta_{n,IPTW}$ as initial estimator. For more details, we refer to van der Laan and Robins (2002) (pages 118-119).

The likelihood-based estimator.

Consider the identifiability result (18) for $DE(a,W) = E(\sum_z Q_0(z \mid W)(Y_{az} - Y_{0z}) \mid W)$ given by:

$$DE(a,W) = \int_z \{E(Y_{az} \mid W) - E(Y_{0z} \mid W)\} Q_0(z \mid W). \qquad (26)$$

The fact that $E_0(DE(a,W) \mid V) = m(a, V \mid \beta_0)$ suggests 1) to estimate $E(Y_{az} \mid W) - E(Y_{0z} \mid W)$ according to a marginal structural model, 2) to obtain a fitted $\hat{DE}(a,W)$ by plugging in this estimator in (26), and 3) to regress the vector $(DE(a,W_i) : a)$, on V according to the repeated measured regression 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), and are therefore not repeated here.

Statistical Inference.

The statistical inference can be presented in precisely the same manner as in Section 4.

References

J. Pearl. Direct and indirect effects. In M. Kaufmann, editor, *Proceedings of the Seventeenth Conference on Uncertaint 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.

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, 2000a.

J.M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, 2000b.

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

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

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