*University of California, Berkeley*

U.C. Berkeley Division of Biostatistics Working Paper Series

*Year* 2011 *Paper* 293

# Identification and Efficient Estimation of the Natural Direct Effect Among the Untreated

Samuel D. Lendle[*] Mark J. van der Laan[†]

[*] University of California, Berkeley, School of Public Health - Division of Biostatistics, lendle@stat.berkeley.edu

[†] University of California, Berkeley; School of Public Health - Division of Biostatistics, laan@berkeley.edu

# Identification and Efficient Estimation of the Natural Direct Effect Among the Untreated

Samuel D. Lendle and Mark J. van der Laan

## Abstract

The natural direct effect (NDE), or the effect of an exposure on an outcome if an intermediate variable was set to the level it would have been in the absence of the exposure, is often of interest to investigators. In general, the statistical parameter associated with the NDE is difficult to estimate in the non-parametric model, particularly when the intermediate variable is continuous or high dimensional. In this paper we introduce a new causal parameter called the natural direct effect among the untreated, discus identifiability assumptions, and show that this new parameter is equivalent to the NDE in a randomized control trial. We also present a targeted minimum loss estimator (TMLE), a locally efficient, double robust substitution estimator for the statistical parameter associated with this causal parameter. The TMLE can be applied to problems with continuous and high dimensional intermediate variables, and can be used to estimate the NDE in a randomized controlled trial with such data. Additionally, we define and discuss the estimation of three related causal parameters: the natural direct effect among the treated, the indirect effect among the untreated and the indirect effect among the treated.

# 1 Introduction

Researchers are often interested in not only the total effect of an exposure on an outcome, but also how the exposure acts to effect the outcome by way of a mediator. For example, suppose there is a dietary intervention designed to reduce the risk of acute myocardial infarction (AMI) which also tends to result in weight loss. An investigator may be interested in the effect of diet on risk of AMI that is not due to weight loss. Specifically, she may ask "how would a patient's risk of AMI have changed due to the intervention diet if their weight had been set to whatever it would have been had the patient not been on the intervention diet?" This sort of effect is known as a natural direct effect (Robins and Greenland, 1992; Pearl, 2001).

Many methods for estimating the natural direct effect require consistent estimation of the conditional distribution of the intermediate variable conditional on treatment and baseline covariates, e.g. Petersen et al. (2006); van der Laan and Petersen (2008); VanderWeele (2009); VanderWeele and Vansteelandt (2010). If the intermediate variable, like weight loss in the example above, is continuous or multivariate, this becomes difficult without relying on strong parametric assumptions. Jo et al. (2011) describe a propensity score based estimation method but it is restricted to settings with a binary mediator.

Tchetgen Tchetgen and Shpitser (2011) and Zheng and van der Laan (2011) develop semiparametric theory for the natural direct effect and present multiply robust estimators for the statistical parameter. Tchetgen Tchetgen and Shpitser (2011) present an estimator based on an estimating equation approach and Zheng and van der Laan (2011) develops a targeted minimum loss estimator. When the distribution of the treatment conditional on baseline covariates is known or consistently estimated, for example in a randomized controlled trial, consistency and efficiency of these estimators only depends on the mediator distributions conditional on baseline covariates and treatment through a ratio. In such a setting, the estimators can be modified to use estimates of the distribution of treatment conditional on baseline covariates and the intermediate variable in place of an estimate of the distribution of the mediator given covariates and treatment (Zheng and van der Laan, 2011).

In this paper, we propose a new causal parameter which we call the natural direct effect among the untreated. We show that this parameter is identifiable under similar conditions to those of the natural direct effect, and in a randomized control trial, it is equal to the natural direct effect. Additionally we present a targeted minimum loss estimator (TMLE) for the statistical parameter. We also define and discuss the estimation of the natural direct effect among the treated as well as the indirect effect among the untreated and among the treated.

# 2 The counterfactual framework and natural direct effects

Following Robins and Greenland (1992) and Pearl (2001), we define natural direct effects using the counterfactual framework. For an individual, let $Z_a$ be the counterfactual value of the intermediate variable, or mediator, had their exposure, $A$, been set to $a$ for all $a \in \mathcal{A}$, the set of all possible exposures. Similarly, let $Y_{az}$ be the counterfactual outcome had the individual's exposure and intermediate been set to $a$ and $z$, respectively, for all $(a, z) \in \mathcal{A} \times \mathcal{Z}$. These values are called counterfactual because in practice, a researcher can only observe the mediator and outcome for the exposure level that an individual was observed to have.

Without loss of generality, let exposure $A = 0$ be the reference or untreated level. The individual natural direct effect is defined as $Y_{aZ_0} - Y_{0Z_0}$. The natural direct effect is also known as the "pure direct effect" (Robins and Greenland, 1992). This is interpreted as the change in outcome due to exposure $a$ had the mediator been set to the level it would have been under exposure 0. Note that this quantity is different than the individual controlled direct effect, $Y_{az} - Y_{0z}$, where the mediator is set to some specific level $z$, not necessarily equal to $Z_0$.

# 3 Identifiability

Similarly to van der Laan and Petersen (2008), we assume there exists a random variable $X := \{W, A, Z_a, Y_{az} : a \in \mathcal{A}, z \in \mathcal{Z}\}$. In addition, we assume $O := \{W, A, Z = Z_A, Y = Y_{AZ}\}$ is a missing data structure on $X$ where $A$ is the observed exposure, and $W$ represents a possibly multivariate baseline covariate. As implied by the definition of $O$, we also assume consistency, that $Z$ is the counterfactual mediator under the observed exposure, and $Y$ is the counterfactual outcome under the observed exposure and mediator.

Let $\mathcal{M}$ be the set of possible probability distributions $P$ for $O$, and call the true distribution of $O$ $P_0$. The set $\mathcal{M}$ is called the statistical model. For sake of presentation suppose $O$ is a discrete random variable, so $P$ represents a probability. To allow for continuous random variables, we can assume $\mathcal{M}$ is dominated by a common measure and define densities with respect to that measure. The likelihood of $O$ can be factorized as

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

A causal parameter is a mapping from the full data model into the real numbers, $\Psi^F : \mathcal{M}^F \to \mathbb{R}^k$, where $\mathcal{M}^F$ is the set of all possible data generating distributions of $X$, known as the causal model or full data model. Let $F_{X_0} \in \mathcal{M}^F$ be the true distribution of $X$. In order to have any hope of estimating the causal parameter $\Psi^F(F_{X0})$ of interest, we must be able to write it as a functional of only the distribution of the observed data $O$. That is, we need make some assumptions on $\mathcal{M}^F$ to be able to find some $\Psi$ such that $\Psi^F(F_X) = \Psi(P(F_X))$ for all $F_X \in \mathcal{M}^F$ where $P(F_X)$ is the distribution of $O$ implied by $F_X$.

**Assumption 1** (Randomization)**.**

$$(A, Z) \perp Y_{az} \mid W$$

*and*

$$A \perp Z_a \mid W$$

Assumption 1 can be interpreted as assuming that the exposure and mediator share no common causes with the outcome and that the exposure shares no common causes with the mediator that are not measured in the set of baseline covariates.

**Assumption 2** (Positivity)**.** *For $a \in \mathcal{A}$, $P_0(A = a \mid Z = z, W = w) > 0$ for all $(z, w)$ where $P_0(Z = z, W = w \mid A = 0) > 0$.*

The positivity assumption is also known as experimental treatment assignment (ETA) assumption, and can be interpreted as assuming for every strata of $W$ and $Z$ that can occur when $A = 0$, treatment level $a$ has a non-zero probability of occurring.

**Assumption 3.**

$$E(Y_{az} - Y_{0z} \mid Z_0 = z, W) = E(Y_{az} - Y_{0z} \mid W)$$

Consider the causal parameter

$$\Psi^F(F_X) = DEU(a) = E\Big\{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A = 0\Big\}, \qquad (1)$$

a generalized natural direct effect among the untreated population.

**Theorem 1.** *(i) Under the randomization assumption (Assumption 1) and the positivity assumption (Assumption 2), $DEU(a)$ is identifiable. (ii) Additionally under Assumption 3, $DEU(a)$ equals the causal parameter $E(Y_{aZ_0} - Y_{0Z_0} \mid A = 0)$.*

Call the causal parameter $E(Y_{aZ_0} - Y_{0Z_0} \mid A = 0)$ the natural direct effect among the untreated, as it is the average of individual natural direct effects among those who have treatment $A = 0$. Theorem 1 is closely related to the identifiability results in van der Laan and Petersen (2004), and the Assumptions 1 to 3 are analogous to the assumptions for identifiability of

$$DE(a) = E\Big\{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W)\Big\},$$

a generalized natural direct effect discussed in van der Laan and Petersen (2008) and for $E(Y_{aZ_0} - Y_{0Z_0})$, the natural direct effect. For other discussions of identifiability of direct effects, see Robins and Greenland (1992); Pearl (2001); Hafeman and Vanderweele (2011); Imai et al. (2010); Pearl (2011).

Assumption 3 means that the expected individual direct effect given $W$ with an intermediate fixed at $z$ does not depend on the counterfactual intermediate value $Z_0$. Because the NDE and the NDE among the untreated depend on the counterfactual value $Y_{aZ_0}$, which can not ever be observed in a real life experiment, it is not surprising that an unintuitive and somewhat strong assumption like Assumption 3 is required for identification. Even when Assumption 3 does not hold, the causal parameter $DEU(a)$ is interpretable as an average of controlled direct effects averaged with respect to the distribution of the counterfactual $Z_0$ conditional on the distribution of baseline covariates among the untreated group.

Under the randomization and positivity assumptions, we know that $DEU(a)$ is identifiable, and we can write $DEU(a)$ as a functional of the observed data generating distribution:

$$\begin{aligned}\Psi(P_0) =& E\Big(\Big[\sum_z \{E(Y \mid A=a, Z=z, W) - E(Y \mid A=0, Z=z, W)\} \qquad (2)\\ &\qquad P(Z=z \mid A=0, W)\Big] \mid A=0\Big).\end{aligned}$$

**Theorem 2.** *(i) If $A$ is completely randomized (i.e. $A \perp (W, Z_a, Y_{az})$), then $DEU(a) = DE(a)$. (ii) Additionally under Assumption 3, $E(Y_{aZ_0} - Y_{0Z_0}|A = 0) = E(Y_{aZ_0} - Y_{0Z_0})$.*

In a randomized control trial (RCT) where subjects are randomly assigned to a treatment $a \in \mathcal{A}$ independent of baseline covariates $W$, the conditions for Theorem 2 (i) are satisfied. Furthermore, $A \perp (W, Z_a, Y_{az})$ implies that $A \perp Y_{az} \mid W$ and $A \perp Z_a \mid W$, so the only randomization assumption which is not automatically satisfied is $Z \perp Y_{az} \mid W$.

## 4 Estimation

In Section 3, we defined the statistical parameter that we are interested in estimating, $\Psi(P_0)$ in (2). Let $B = (W, Z)$ and without loss of generality let the exposure level of interest $a = 1$. The target statistical parameter can be written as

$$\psi_0 = \Psi(P_0) = E_0\{E_0(Y \mid A=1, B) - E_0(Y \mid A=0, B) \mid A=0\} \qquad (3)$$

Under other causal models, $\Psi(P_0)$ can be interpreted as other interesting causal parameters. For example, if $B = Z$ and Assumptions 1 to 3 are strengthened to $(A, Z) \perp Y_{az}$, $A \perp Z_a$, $P_0(A=1 \mid Z=z, W=w) \in (0,1)$ almost everywhere, and $E(Y_{az} - Y_{0z} \mid Z_0 = z) = E(Y_{az} - Y_{0z})$, then $\Psi(P_0)$ is the natural direct effect as defined in van der Laan and Rose (2011, Chapter 8). Under a different causal model, $\Psi(P_0)$ is equivalent to the so called average treatment effect among the untreated (van der Laan, 2010; Hahn, 1998).

The functional $\Psi$ is a mapping from the non-parametric statistical model $\mathcal{M}$ to $\mathbb{R}$. For a distribution $P \in \mathcal{M}$, let $\bar{Q}(a, b) = E_P(Y \mid A=a, B=b)$, $g(a \mid b) = P(A=a \mid B=b)$, and $Q_B(b) = P(B=b)$ for $a \in \{0,1\}$ and $b \in \mathcal{B}$, the support of $B$. Let the subscript 0 denote the truth and the subscript $n$ denote an estimate based on $n$ independent observations $O_i = (B_i, A_i, Y_i)$ for $i = 1, \ldots, n$. For example, $\bar{Q}_0$ is the true conditional mean of $Y$ and $\bar{Q}_n$ is an estimate. The mapping $\Psi$ only depends on $P$ through $Q = (\bar{Q}, Q_B)$ and $g$, so with an abuse of notation we write the parameter as

$$\Psi(Q, g) = \sum_b \left[ \{\bar{Q}(1, b) - \bar{Q}(0, b)\} \frac{g(0|b)Q_B(b)}{\sum_b \{g(0|b)Q_B(b)\}} \right]$$

Bickel et al. (1993) show that a regular estimator for a statistical parameter in a semiparametric model is asymptotically efficient, (i.e. the estimator has minimal asymptotic variance,) if it is asymptotically linear with influence curve (influence function) equal to the efficient influence curve. This minimal asymptotic variance is known as the semiparametric efficiency bound and is the variance of the efficient influence curve. The non-parametric model $\mathcal{M}$ is a special case of a semiparametric model where there are no restrictions on the possible distributions of $O$. The efficient influence curve for $\Psi$ at $P \in \mathcal{M}$, derived in van der Laan (2010), is

$$\begin{aligned} D^*(P) = D^*(Q, g, \Psi(Q, g)) = & \left\{ \frac{I(A=1)}{P(A=0)} \frac{g(0 \mid B)}{g(1 \mid B)} - \frac{I(A=0)}{P(A=0)} \right\} \{Y - \bar{Q}(A, B)\} \\ & + \frac{I(A=0)}{P(A=0)} \{\bar{Q}(1, B) - \bar{Q}(0, B) - \Psi(Q, g)\} \end{aligned}$$

where $I(\cdot)$ is an indicator function. The semiparametric efficiency bound for an analogous statistical parameter, where the difference is conditioned on $A = 1$, is also derived in Hahn (1998).

The efficient influence curve for $\Psi(P_0)$ has the double robustness property. That is,

$$P_0 D^*(Q, g_0, \psi_0) = P_0 D^*(Q_0, g, \psi_0) = 0,$$

where $Pf := \int f(o) dP(o) = \sum_o f(o) P(O = o)$ is the expectation of $f$ under distribution $P$. This means that if we have an estimator that solves the efficient influence curve equation, (i.e. $P_n D^*(Q_n, g_n, \Psi(Q_n, g_n)) = 0$,) then it is consistent if at least one of the estimators $Q_n$ or $g_n$ are consistent for $Q_0$ or $g_0$ under regularity conditions (van der Laan, 2010). Additionally, the efficiency bound is achieved if both $Q_n$ and $g_n$ are consistent estimators for $Q_0$ and $g_0$, so such an estimate is locally efficient at $P_0$.

In van der Laan and Rose (2011, Chapter 8) and van der Laan (2010), a targeted minimum loss estimator (TMLE) is developed for $\Psi(P_0)$. The TMLE solves the efficient influence curve and is a locally efficient, double robust estimator. It is also a substitution or plug-in estimator in the sense that estimators for $Q_0$ and $g_0$ can be plugged into the mapping $\Psi$ to calculate an estimate as

$$\Psi(Q_n, g_n) = \frac{1}{\sum_{i=1}^n I(A_i = 0)} \sum_{i=1}^n I(A_i = 0) \{\bar{Q}_n(1, B_i) - \bar{Q}_n(0, B_i)\} \qquad (4)$$

for some estimates $Q_n$ and $g_n$. That is, the estimate is the difference $\bar{Q}_n(1, B_i) - \bar{Q}_n(0, B_i)$ for each individual averaged with respect to the empirical distribution of $B$ given $A = 0$. We review the TMLE for $\Psi(P_0)$ here.

Begin by constructing initial estimates for $\bar{Q}_0$ and $g_0$ called $\bar{Q}_n^0$ and $g_n^0$. If we have expert background knowledge about the functional forms of $\bar{Q}_0$ and $g_0$, they can be estimated by parametric models. In general there is not enough background knowledge to support parametric models, and $\bar{Q}_n^0$ and $g_n^0$ should be constructed by some non-parametric data adaptive learning algorithm such as the super learner (van der Laan et al., 2007), which combines machine learning algorithms and parametric models using cross validation. To calculate the TMLE, we update the initial estimates $\bar{Q}_n^0$ and $g_n^0$ to $\bar{Q}_n^*$ and $g_n^*$, and then plug them in to $\Psi$, so the final estimate is $\Psi(Q_n^*, g_n^*)$, where $Q_n^* = (\bar{Q}_n^*, Q_{Bn})$ and $Q_{Bn}$ is the empirical distribution of $B$.

For now, suppose $Y$ is either binary or bounded between 0 and 1. If it is not, the following algorithm can be modified slightly as explained below. To update the initial estimates, for $j = 1, 2, \ldots$, calculate until convergence

$$\operatorname{logit} \bar{Q}_n^j(A, B) = \operatorname{logit} \bar{Q}_n^{j-1}(A, B) + \epsilon_{1n}^j C_1^{j-1}(A, B) \tag{5}$$

and

$$\operatorname{logit} g_n^j(0 \mid B) = \operatorname{logit} g_n^{j-1} + \epsilon_{2n}^j C_2^{j-1}(B)$$

where

$$C_1^{j-1}(A, B) = \left\{ \frac{I(A = 1)}{P_n(A = 0)} \frac{g_n^{j-1}(0 \mid B)}{g_n^{j-1}(1 \mid B)} - \frac{I(A = 0)}{P_n(A = 0)} \right\},$$

$$C_2^{j-1}(B) = \frac{1}{P_n(A = 0)} \{ \bar{Q}_n^{j-1}(1, B) - \bar{Q}_n^{j-1}(0, B) - \Psi(Q_n^{j-1}, g_n^{j-1}) \},$$

and $P_n(A = 0)$ is the empirical proportion of observations with $A = 0$. The coefficients $\epsilon_{1n}^j$ and $\epsilon_{2n}^j$ in the above expressions are the maximum likelihood estimates in the logistic regression models

$$\operatorname{logit} \bar{Q}(A, B) = \epsilon_1^j C_1^{j-1}(A, B) + \operatorname{logit} \bar{Q}_n^{j-1}(A, B)$$

and

$$\operatorname{logit} g(0 \mid W) = \epsilon_2^j C_2^{j-1} + \operatorname{logit} g_n^{j-1}(0 \mid B).$$

These coefficients can be calculated with standard software where $\bar{Q}_n^{j-1}(A, B)$ and $g_n^{j-1}(0 \mid B)$ are offset terms. Convergence is reached when both $\epsilon_{1_n}^j$ and $\epsilon_{2n}^j$ are close to 0 and so estimates of $\bar{Q}_0$ and $g_0$ are changing very little. Set $\bar{Q}_n^* = \bar{Q}_n^j$ and $g_n^* = g_n^j$ at the last iteration.

If $Y$ is not bounded by 0 and 1, the updating steps for $\bar{Q}_n$ can be altered slightly in one of two possible ways. The first way is the simplest. Instead of updating the estimate $Q_n^i$ on the logit scale, we can updated it on the linear scale by replacing (5) with

$$\bar{Q}_n^j(A, B) = \bar{Q}_n^{j-1}(A, B) + \epsilon_{1n}^j C_1^{j-1}(A, B)$$

where $\epsilon_{1n}^j$ is estimated with maximum likelihood or least squares in the linear model

$$\bar{Q}(A,B) = \epsilon_1^j C_1^{j-1}(A,B) + \bar{Q}_n^{j-1}(A,B).$$

In small samples, when $Y$ is continuous and bounded by $l$ and $u$ with $l < u$, this linear fluctuation can yield final estimates that do not respect the bound of the model. For example, suppose $Y$ is a percentage between 0 and 100. A linear fluctuation could potentially yield estimates less than $-100$ or greater than 100.

The second modification to the algorithm avoids this situation by transforming $Y$ to $Y'$ bounded between 0 and 1. After estimating $Q_n^0(A,B)$, calculate $Y' = (Y-l)/(u-l)$ and $Q_n'^0 = (Q_n^0(A,B)-l)/(u-l)$, and perform the updating steps with $Y'$ and $Q_n'^0$ in place of $Y$ and $Q_n^0$. After convergence, calculate the final estimate by multiplying $\Psi(Q_n^*, g_n^*)$ by $u-l$.

In order to conduct hypothesis tests and construct confidence intervals, we need to approximate the sampling distribution of the TMLE $\Psi(Q_n^*, g_n^*)$. Under regularity conditions on the initial estimates $Q_n^0$ and $g_n^0$, the TMLE is regular and asymptotically linear (van der Laan and Rose, 2011), so $\sqrt{n}(\Psi(P_n^*) - \Psi(P_0)) \xrightarrow{d} N(0,\sigma^2)$. When $Q_n^0$ and $g_n^0$ are consistent estimators for $Q_0$ and $g_0$, the variance $\sigma^2$ is the variance of the efficient influence curve. In order to estimate the variance $\sigma^2$, we can calculate an estimate of the influence of each observation by plugging $O_i$ into the estimated influence curve $D^*(P_n^*)$ of the distribution $P_n^* = (Q_n^*, g_n^*)$, and calculate the sample variance of these influences. Wald type tests can be performed, and confidence intervals can be constructed with the estimated variance $\sigma_n^2$.

When either $Q_n^0$ or $g_n^0$ is not consistent, the influence curve based variance estimate is biased and not guaranteed to be conservative. If one assumes $g_n^0$ is a consistent MLE, then one can compute a correction term for the influence curve which only depends on the behavior of $g_n^0$ (van der Laan and Robins, 2003, Section 2.3.7) Alternatively, the non-parametric bootstrap can be used to estimate the variance of the TMLE in the standard way by resampling $n$ observations many times from the original data and calculating the TMLE for each resampled $n$ observations. The variance is estimated as the sample variances of the estimates of $\Psi(P_0)$ from each resampled data set. When initial estimates $Q_n^0$ and $g_n^0$ are differentiable functionals of the empirical distribution, as is the case for parametric maximum likelihood estimators, then the TMLE is also differentiable. In this case we know the bootstrap estimate of the variance is consistent (Gill et al., 1989).

# 5 Simulation studies

To explore the performance of the TMLE in Section 4 we compare the TMLE to other types of estimators in a simulation studies based on two data generating distributions. The first alternative estimator is known as the G-computation or maximum likelihood based estimator (MLE), and depends only on an initial estimate $\bar{Q}_n^0$. The estimate is computed by plugging $\bar{Q}_n^0$ into (4) and averaging with respect to the empirical distribution of $B$ where $A = 0$. An inverse probability of treatment weighted (IPTW) type estimator is also presented, which is a function of an initial estimate of $g_0$. The estimate is computed as

$$\psi_n = n^{-1} \sum_i \left\{ \frac{I(A_i = 1)}{P_n(A_i = 0)} \frac{g_n^0(0 \mid B_i)}{g_n^0(1 \mid B_i)} - \frac{I(A_i = 0)}{P_n(A_i = 0)} \right\} Y_i.$$

See Robins et al. (2000) for a detailed treatment of IPTW estimators. Because these two estimates depend only on either $Q$ or $g$, they are not double robust and we expect them to be biased if estimates of $Q_0$ or $g_0$ are not consistent.

For the first data generating distribution, suppose we observe two baseline independent baseline covariates. The first, $W_1$ has a Bernoulli distribution with mean 0.3, and the second, $W_2$ has a standard normal distribution. We also observe a binary treatment variable $A$, and a mediator $Z$. Suppose $Z$ has a normal distribution with mean $|3W_1|$ and variance one, and $A$ equals one with probability $\text{logit}^{-1}(-2.5 + 3W_1 + 0.2Z)$. Also suppose we observe a binary outcome, $Y$, which is one with probability $\text{logit}^{-1}(1.4A - 2.5Z + W1)$. Call the true distribution of $O = \{W_1, W_2, A, Z, Y\}$ $P_0$.

The statistical parameter $\psi(P_0) \approx 0.0872$ and the variance bound for a sample of size $n$ is approximately $1.004/n$. The true parameter and variance bound were computed by Monte Carlo simulation. By the construction of $P_0$ we can see that the true $\bar{Q}_0$ is contained in a main terms logistic regression model including $W_1$, $W_2$, $A$, and $Z$ as explanatory variables, and the true $g_0$ is contained in a main terms logistic regression model including $W_1$, $W_2$, and $Z$ as explanatory variables. For sake of illustration, we construct initial estimates $\bar{Q}_n^0$ and $g_n^0$ using logistic regression, which we know will be consistent as long as all necessary independent variables are included in the model. In practice we would turn to data adaptive methods for the initial estimates when we do not have enough knowledge to guarantee that estimators based on parametric models are consentent for $\bar{Q}_0$ and $g_0$. In the simulations, the misspecified model for $\bar{Q}$ is a main terms logistic regression model with only $A$ as an explanatory variable, and the misspecified model for $g$ has only $Z$ as an explanatory variable.

Results from 1,000 dataset drawn from $P_0$ of size $n = 50$, $n = 200$ and $n = 1000$ are shown in Table 1. When the models are correctly specified, all three estimators have low bias, and the variance of TMLE estimates approaches the efficiency bound as sample size increases, demonstrating that the TMLE is locally efficient. We also see that bootstrap estimates of the variance are close to the observed variance. When the model for $\bar{Q}_0$ is misspecified, we see the MLE has a large bias which does not decrease with sample size. Similarly when the model for $g_0$ is misspecified, the IPTW estimator has a large bias. However, when one of the models for $\bar{Q}_0$ or $g_0$ is misspecified, TMLE still has low bias, demonstrating the double robustness property.

In a second example, suppose $W_1$ and $W_2$ are distributed as above. Suppose $A$ is completely randomized as in an RCT, and is one with probability 0.25, so there are three observation with $A = 0$ for each observation with $A = 1$. Suppose $Z$ is has a normal distribution with mean $3 + 2A + W_2$, and variance one, and

Table 1: Simulation results from an observational study. Variance bounds were 0.0201, 0.005, and 0.001 for sample sizes 50, 200, and 1000 respectively. Sample sizes are in parentheses.

| Model | Bias | | | Observed Variance | | | Bootstrap Var. Est. | | |
|---|---|---|---|---|---|---|---|---|---|
| | (50) | (200) | (1000) | (50) | (200) | (1000) | (50) | (200) | (1000) |
| $Q$, $g$ correct | | | | | | | | | |
| TMLE | $-0.007$ | $-0.005$ | 0.002 | 0.029 | 0.007 | 0.001 | 0.048 | 0.009 | 0.001 |
| MLE | 0.010 | 0.000 | 0.002 | 0.024 | 0.003 | 0.001 | 0.031 | 0.003 | 0.001 |
| IPTW | $-0.015$ | $-0.003$ | 0.001 | 0.088 | 0.015 | 0.003 | $4e+24$ | 0.017 | 0.003 |
| $Q$ misspecified | | | | | | | | | |
| TMLE | $-0.021$ | $-0.013$ | $-0.001$ | 0.049 | 0.011 | 0.002 | 0.055 | 0.010 | 0.002 |
| MLE | $-0.024$ | $-0.025$ | $-0.023$ | 0.014 | 0.004 | 0.001 | 0.015 | 0.004 | 0.001 |
| $g$ misspecified | | | | | | | | | |
| TMLE | 0.004 | 0.000 | 0.002 | 0.022 | 0.003 | 0.001 | 0.029 | 0.004 | 0.001 |
| IPTW | 0.041 | 0.044 | 0.045 | 0.018 | 0.004 | 0.001 | 0.027 | 0.004 | 0.001 |

$Y$ is binary and equals one with probability $\text{logit}^{-1}(-1+.4A+W_2+.5Z)$. Now let $O=\{W_1,W_2,A,Z,Y\}$ be distributed as $P_0'$.

For this distribution $\Psi(P_0')\approx 0.0656$. Although $A$ is independent of $W_1$ and $W_2$, $g_0'(1\mid W_1,W_2,Z)$ can be very close to zero for small values of $Z$. This is a near positivity violation, and therefore the variance of $D^*(P_0')$ is large, and the variance bound for a sample of $n$ observations is approximately $35.48/n$.

Results from $1,000$ datasets drawn from $P_0'$ are shown in Table 2. We observe that the IPTW estimator has much higher bias and variance relative to the other estimators, because the estimator weights by $1/g_n(1\mid W_1,W_2,Z)$ which can be very large when the positivity assumption is violated. Although TMLE relies on an estimate of $g_0$, the observations with small $g_n(1\mid B_i)$ cannot affect the estimate too much when the logistic fluctuation is used, and the TMLE performs well relative to the IPTW estimator even in low sample sizes. When both models for $Q_0$ and $g_0$ are correctly specified, the TMLE has variance lower than the efficiency bound. This is not unusual for small sample sizes but, because bias is so low, this indicates that the asymptotic properties of the estimator may not have fully taken effect for $n=200$ or even 1000. This is likely also due to the positivity violation. The MLE also performs well as it does not rely on an estimate of $g_0$ at all. This can be seen as an advantage of substitution estimators. The double robustness of the TMLE is demonstrated again here, where bias generally decreases as sample size increases, even when one of the models for $Q_0$ or $g_0$ is misspecified.

Table 2: Simulation results from an RCT with positivity violations. Variance bounds were 0.7095, 0.1774, and 0.0355 for sample sizes 50, 200, and 1000 respectively. Sample sizes are in parentheses.

| Model | Bias | | | Observed Variance | | | Bootstrap Var. Est. | | |
|---|---|---|---|---|---|---|---|---|---|
| | (50) | (200) | (1000) | (50) | (200) | (1000) | (50) | (200) | (1000) |
| $Q$, $g$ correct | | | | | | | | | |
| TMLE | 0.017 | $-0.002$ | 0.001 | 0.082 | 0.055 | 0.019 | 0.070 | 0.035 | 0.014 |
| MLE | 0.010 | $-0.003$ | $-0.001$ | 0.039 | 0.009 | 0.002 | 0.038 | 0.008 | 0.002 |
| IPTW | $-0.178$ | $-0.058$ | $-0.015$ | 0.302 | 0.201 | 0.059 | $5e+25$ | 1.536 | 0.063 |
| $Q$ misspecified | | | | | | | | | |
| TMLE | $-0.006$ | $-0.043$ | $-0.043$ | 0.107 | 0.064 | 0.020 | 0.080 | 0.045 | 0.016 |
| MLE | 0.146 | 0.139 | 0.140 | 0.020 | 0.005 | 0.001 | 0.020 | 0.005 | 0.001 |
| $g$ misspecified | | | | | | | | | |
| TMLE | 0.000 | $-0.016$ | $-0.009$ | 0.057 | 0.024 | 0.006 | 0.048 | 0.017 | 0.005 |
| IPTW | $-0.186$ | $-0.172$ | $-0.169$ | 0.063 | 0.022 | 0.004 | 0.084 | 0.024 | 0.004 |

## 6 Discussion

In this manuscript we proposed a new causal parameter called the natural direct effect among the untreated, and we provide identifiability results in Section 3. In Section 4, we describe a targeted minimum loss estimator that is a locally efficient and double robust substitution estimator for the statistical parameter $\Psi(P_0)$. In Theorem 2 we show when $A$ is completely randomized, such as in an RCT, this natural direct effect among the untreated is equal to the natural direct effect, and therefore the natural direct effect can be estimated with the method in Section 4. Even when $A$ is not completely randomized, an estimate of $\Psi(P_0)$ can always be interpreted as the $DEU(a)$ under the assumptions in Section 3, that is, an average of direct effects weighted by the empirical distribution of baseline covariates $W$ among the unexposed subjects with $A = 0$.

We point out that efficient estimators for $\Psi(P_0)$ in the non-parametric model are not fully efficient in the semiparametric model where $A$ is completely randomized. When the knowledge that $A \perp W$ is ignored and $g_0(A \mid B) = P_0(A \mid W, Z)$ is estimated without restriction, some information about $\Psi(P_0)$ is lost and the efficient influence curve in this semiparametric model is not equal to $D^*$ (Tchetgen Tchetgen and Shpitser, 2011; Zheng and van der Laan, 2011). Although the TMLE in Section 4 is not fully efficient when $A$ is completely randomized, we argue it is still useful as an alternative and relatively simple estimator for the NDE in addition to being an estimator for the NDE among the untreated. Below we discuss other causal parameters to which the TMLE can be applied.

In addition to the NDE and the NDE among the untreated, researchers may also be interested in the NDE among the treated, defined as $E(Y_{aZ_0} - Y_{0Z_0} \mid A = a)$. Under appropriate identifiability conditions, this causal parameter corresponds to the statistical parameter

$$\Psi'(P_0) = \quad E\Big(\Big[\textstyle\sum_z \{E(Y \mid A=a, Z=z, W) - E(Y \mid A=0, Z=z, W)\} \\ P(Z=z \mid A=0, W)\Big] \mid A=a\Big).$$

Because the conditional probability of $Z$ is conditional on $A = 0$ inside the square brackets, but the expectation of the expression in square brackets is conditioned on $A = a$, $\Psi'(P_0)$ cannot be written in the form of (3) and cannot be estimated using a method similar to that in Section 4. However, when there are only two levels of treatment so $A$ is binary, then $\Psi^*(P_0) = \Psi'(P_0)P_0(A = 1) + \Psi(P_0)P_0(A = 0)$ where

$$\Psi^*(P_0) = \quad E\Big(\Big[\textstyle\sum_z \{E(Y \mid A=1, Z=z, W) - E(Y \mid A=0, Z=z, W)\} \\ P(Z=z \mid A=0, W)\Big]\Big)$$

is the statistical parameter associated with the natural direct effect. We can write $\Psi'(P_0) = \{\Psi^*(P_0) - \Psi(P_0)P_0(A = 0)\}/P_0(A = 1)$. Based on this we can see that $\Psi'(P_0)$ can be estimated using an estimate for $\Psi^*(P_0)$ such as those proposed by Zheng and van der Laan (2011) and Tchetgen Tchetgen and Shpitser (2011) as well as an estimate for $\Psi(P_0)$ based on the methodology in Section 4.

Another causal parameter that may be of interest to researchers is called the indirect effect (IE) among the untreated, defined as $E(Y_{aZ_a} - Y_{aZ_0} \mid A = 0)$. This definition is analogous to the total indirect effect of Robins and Greenland (1992) and the indirect effect of van der Laan and Petersen (2004). Similarly to the total effect (TE), the TE among the untreated or average effect of treatment among the untreated (ATU), defined as $E(Y_{aZ_a} - Y_{0Z_0} \mid A = 0)$ in current notation, can be decomposed as the sum of the NDE among the untreated and the IE among the untreated. That is,

$$E(Y_{aZ_a} - Y_{0Z_0} \mid A=0) = E(Y_{aZ_a} - Y_{aZ_0} \mid A=0) + E(Y_{aZ_0} - Y_{0Z_0} \mid A=0).$$

Because of this decomposition, if the ATU and the NDE among the untreated are identifiable, the IE among the untreated can also be identified and can be estimated based on estimates of the ATU and the NDE among the untreated. Identifiability of the average treatment effect among the (un)treated is discussed in van der Laan (2010). Analogously, this relationship also holds for the TE among the treated, the NDE among the treated, and the IE among the treated so the indirect effect among the untreated can be estimated similarly.

A final alternative causal parameter of interest may be defined as $E(Y_{aZ_a} - Y_{0Z_a} \mid A = a)$. This parameter is similar to the NDE among the treated, but the intermediate variable is set to the value it would have been under treatment $a$ instead of treatment 0. Under appropriate identifiability assumptions, this is equal to the statistical parameter

$$E_0\{E_0(Y \mid A=1, B) - E_0(Y \mid A=0, B) \mid A=a\}. \qquad (6)$$

This statistical parameter is similar to (3), but now the difference is conditional on $A = a$. An analogous estimator to that developed in Section 4 could be used for this parameter, or one could simply code a new treatment variable $A' = 0$ when $A = a$ and $A' = a$ when $A = 0$, and implement the TMLE described above. Multiplying this TMLE by negative one yields an estimate for (6).

## Acknowledgements

The authors thank Wenjing Zheng helpful discussion. Mark van der Laan was supported by NIH grant R01 AI074345-04.

## References

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press, Baltimore, 1993. ISBN 0801845416.

R.D. Gill, J.A. Wellner, and J. Præ stgaard. Non-and semi-parametric maximum likelihood estimators and the von Mises method (part 1). *Scandinavian Journal of Statistics*, pages 97–128, 1989. URL `http://www.jstor.org/stable/4616127`.

Danella M Hafeman and Tyler J Vanderweele. Alternative assumptions for the identification of direct and indirect effects. *Epidemiology (Cambridge, Mass.)*, 22(6):753–764, November 2011. ISSN 1531-5487. doi: 10.1097/EDE.0b013e3181c311b2.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2): 315–331, March 1998. ISSN 0012-9682. doi: 10.2307/2998560. URL `http://www.jstor.org/stable/2998560`.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1): 51–71, February 2010. ISSN 0883-4237. doi: 10.1214/10-STS321. URL `http://projecteuclid.org/euclid.ss/1280841733`.

Booil Jo, E.A. Stuart, D.P. MacKinnon, and A.D. Vinokur. The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, 46(3): 425–452, 2011.

Judea Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann, 2001.

Judea Pearl. The mediation formula: a guide to the assessment of causal pathways in nonlinear models. Technical Report July, UCLA, 2011. URL `http://escholarship.org/uc/item/0hz9x8pc.pdf`.

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, M a Hernán, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–60, September 2000. ISSN 1044-3983. URL `http://www.ncbi.nlm.nih.gov/pubmed/10955408`.

J.M. Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, March 1992. ISSN 1044-3983. URL `http://www.jstor.org/stable/3702894`.

Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis : efficiency bounds, multiple robustness, and sensitivity analysis. Working Paper 130, Harvard University Biostatistics Working Paper Series, 2011. URL `http://www.bepress.com/harvardbiostat/paper130`.

Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York, 2003. ISBN 0387955569.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011. ISBN 1441997814.

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), January 2007. ISSN 1544-6115. doi: 10.2202/1544-6115.1309. URL `http://www.ncbi.nlm.nih.gov/pubmed/17910531`.

M.J. van der Laan. Estimation of causal effects of community based interventions. Working Paper 268, U.C. Berkeley Division of Biostatistics Working Paper Series, 2010. URL `http://www.bepress.com/ucbbiostat/paper268/`.

M.J. van der Laan and M.L. Petersen. Estimation of direct and indirect causal effects in longitudinal studies. Working Paper 155, U.C. Berkeley Division of Biostatistics Working Paper Series, 2004. URL `http://www.bepress.com/ucbbiostat/paper155/`.

M.J. van der Laan and M.L. Petersen. Direct effect models. *International Journal of Biostatistics*, 4(1), 2008. doi: 10.2202/1557-4679.1064. URL `http://www.bepress.com/ijb/vol4/iss1/23`.

Tyler J VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26, January 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19234398`.

Tyler J VanderWeele and Stijn Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology*, 172(12):1339–48, December 2010. ISSN 1476-6256.

Wenjing Zheng and M.J. van der Laan. Targeted maximum likelihood estimation of natural direct effect. Working Paper 288, U.C. Berkeley Division of Biostatistics Working Paper Series, 2011. URL `http://www.bepress.com/ucbbiostat/paper288/`.

# A

## A.1 Proofs of theorems

*Proof of Theorem 1*

For (i), by the randomization assumption we can write

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

so

$$\begin{aligned} DEU(a) &= E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A=0\} \\ &= E[E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z \mid W) \mid A=0, W\} \mid A=0] \\ &= E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid A=0, W) P(Z_0 = z \mid W) \mid A=0\} \\ &= E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid W) P(Z_0 = z \mid W) \mid A=0\} \\ &= E([\textstyle\sum_z \{E(Y \mid A=a, Z=z, W) - E(Y \mid A=0, Z=z, W)\} \\ &\qquad P(Z=z \mid A=0, W)] \mid A=0), \end{aligned}$$

therefore $DEU(a)$ is identifiable. For (ii),

$$E(Y_{aZ_0} - Y_{0Z_0} \mid A=0) = E[E\{E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0, A=0, W) \mid A=0, W\} \mid A=0]$$

and

$$\begin{aligned} & E\{E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0, A=0, W) \mid A=0, W\} \\ = \ & \textstyle\sum_z E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0 = z, A=0, W) P(Z_0 = z \mid A=0, W) \\ = \ & \textstyle\sum_z E(Y_{aZ_0} - Y_{0Z_0} \mid Z_0 = z, W) P(Z_0 = z \mid W) \text{ by Assumption 1} \\ = \ & \textstyle\sum_z E(Y_{az} - Y_{0z} \mid Z_0 = z, W) P(Z_0 = z \mid W) \\ = \ & \textstyle\sum_z E(Y_{az} - Y_{0z} \mid W) P(Z_0 = z \mid W) \text{ by Assumption 3} \end{aligned}$$

so

$$\begin{aligned} E(Y_{aZ_0} - Y_{0Z_0} \mid A=0) &= E\{\textstyle\sum_z E(Y_{az} - Y_{0z} \mid W) P(Z_0 = z \mid W) \mid A=0\} \\ &= DEU(a) \qquad \square \end{aligned}$$

*Proof of Theorem 2*

For (i),

$$\begin{aligned} DEU(a) &= E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W) \mid A=0\} \\ &= \textstyle\sum_w \{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W)\} P(W=w|A=0) \\ &= \textstyle\sum_w \{\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W)\} P(W=w) \text{ by } A \text{ completely randomized} \\ &= E\{\textstyle\sum_z (Y_{az} - Y_{0z}) P(Z_0 = z|W)\} \end{aligned}$$

$$= DE(a)$$

The proof for (ii) follows from (i) of this theorem and the proof of Theorem 1 (ii). □