Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Identification, Semiparametric Efficiency, and Quadruply Robust Estimation in Mediation Analysis with Treatment-Induced Confounding

Fan Xia & Kwun Chuen Gary Chan

To cite this article: Fan Xia & Kwun Chuen Gary Chan (2021): Identification, Semiparametric Efficiency, and Quadruply Robust Estimation in Mediation Analysis with Treatment-Induced Confounding, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1990765

To link to this article: https://doi.org/10.1080/01621459.2021.1990765

Identification, Semiparametric Efficiency, and Quadruply Robust Estimation in Mediation Analysis with Treatment-Induced Confounding

Fan Xia
Department of Epidemiology, University of Washington
and
Kwun Chuen Gary Chan
Department of Biostatistics, University of Washington

acadfanxia@gmail.com

Author's Footnote: Fan Xia is UW Data Science Postdoctoral Fellow, National Alzheimer's Coordinating Center, Department of Epidemiology, University of Washington (email: fanxia@uw.edu). Kwun Chuen Gary Chan is Professor, Department of Biostatistics, University of Washington (email: kcgchan@uw.edu).

Abstract

Natural mediation effects are often of interest when the goal is to understand a causal mechanism. However, most existing methods and their identification assumptions preclude treatment-induced confounders often present in practice. To address this fundamental limitation, we provide a set of assumptions that identify the natural direct effect in the presence of treatment-induced confounders. Even when some of those assumptions are violated, the estimand still has an interventional direct effect interpretation. We derive the semiparametric efficiency bound for the estimand, which unlike usual expressions, contains conditional densities that are variational dependent. We consider a reparameterization and propose a quadruply robust estimator that remains consistent under four types of possible misspecification and is also locally semiparametric efficient. We use simulation studies to demonstrate the

proposed method and study an application to the 2017 Natality data to investigate the effect of prenatal care on preterm birth mediated by preeclampsia with smoking status during pregnancy being a potential treatment-induced confounder. Supplementary materials for the article are available online.

Keywords: Copula; Natural Direct Effect; Treatment-induced Confounding; Multiply Robust Estimator; Interventional Direct Effect.

1 Introduction

When a treatment has an aggregated effect on an outcome, its causal mechanism related to intermediate variables along the causal pathway is often of interest. The study of such a treatment effect mechanism involves the estimation of direct and indirect effects. An effect is called direct when it does not act through some intermediate variables, known as mediators. Conversely, the effect that acts through the mediators is called the indirect effect. Intuitively, to evaluate a direct effect, the mediators need to be somehow fixed. Depending on the scientific question of interest, a variety of direct effects can be defined through different ways of fixing the mediators. The natural (pure) effects are most relevant in studying treatment effect mechanisms. The natural direct effect compares potential outcomes under treatment and control conditions with mediators fixed to the value they would have taken had there not been any treatment. The natural indirect effect is thereby defined by subtracting the natural direct effect from the total treatment effect.

Under the potential outcomes framework (Rubin, 1974; Rubin, 1978), the definition of the natural direct effect is formalized by Robins and Greenland (1992) and Pearl (2001). A considerable number of methods have been developed for the identification and inference of the natural direct effect when all confounders are pre-treatment variables (Pearl, 2001; Pearl, 2009; Petersen et al., 2006; Imai et al., 2010; Hafeman and VanderWeele, 2011; Tchetgen Tchetgen and Shpitser, 2012). Although methods for handling pre-

treatment confounders are well studied, they cannot be directly applied to treatment-induced confounders, since they are in the causal pathway between the exposure and outcome of interest. In fact, treatment-induced confounding presents unique challenges to the estimation of natural direct effects, since common identification assumptions in the absence of treatment-induced confounding, such as sequential ignorability (Imai et al., 2010), can no longer identify the natural mediation effects when treatment-induced confounding is present (Avin et al., 2005; VanderWeele and Vansteelandt, 2009). Even under stronger assumptions such as nonparametric structural equation models with independent errors (NPSEM-IE) common in graphical modeling, the natural direct effect is still non-identified in the presence of treatment-induced confounder (Robins and Richardson, 2010; Tchetgen Tchetgen and VanderWeele, 2014). Therefore, both the identification assumptions and estimation of natural mediation effects in the presence of treatment-induced confounding are substantially different from the case with only pre-treatment covariates.

In practice, treatment-induced confounders often exist, particularly when the mediators occur much later than the treatment. In this case, some immediate prognostic factors affected by the treatment can be related to both the mediator and the outcome (Robins, 1999). One example that was given by Vansteelandt and VanderWeele (2012) who considered the effect of adequate prenatal care on preterm birth that mediates through preeclampsia. On one hand, smoking status during pregnancy confounds the relationship between preeclampsia and preterm birth because it reduces the risk of preeclampsia while increasing the likelihood of preterm birth. On the other hand, adequate prenatal care may decrease or eliminate smoking. Therefore, smoking status during pregnancy is a potential treatment-induced confounder between the mediator preeclampsia and the outcome preterm birth. While maternal smoking and preeclampsia can be grouped together as a vector mediator in a mediation analysis (VanderWeele et al., 2014), the joint indirect effect would involve both social and biological

mechanisms which may not have the most desired scientific interpretation. When we are interested in a specific mediator, and the goal is to estimate the direct effect and the indirect effect with respect to this specific mediator, the part of treatment effect that goes through the treatment-induced confounders is not of interest, thus the joint effect is not sufficient to answer the question of interest. Therefore, in many settings, it is often desirable to study the natural direct and indirect effects defined with respect to a specific mediator, while other intermediate variables are treated as confounders.

There has been limited methodological development addressing mediation analysis with treatment-induced confounding, Robins (2003) provides an independence assumption between individual level potential outcome difference and mediators but Petersen et al. (2006) and Imai and Yamamoto (2013) suggest that this assumption is unlikely to hold in practice. Robins and Richardson (2010) and Tchetgen Tchetgen and VanderWeele (2014) each provide additional assumptions to those imposed by structural equation models. Alternatively, estimands different from the natural direct effect are also considered to quantify certain direct and indirect effects in the presence of treatment-induced confounding. VanderWeele et al. (2014) summarize three such approaches to decompose the effect of a treatment when there exists treatment-induced confounding: joint effect of mediators and other treatment-induced confounders (treating the latter as part of the mediators), path-specific effects, and interventional effects.

Avin et al. (2005) and Shpitser (2013) provide identification conditions for path-specific effects. Miles et al. (2020) provide semiparametric inference of a path-specific effect that goes through a mediator without going through its treatment-induced confounders. The interventional direct effect (VanderWeele et al., 2014) is an analog of the natural direct effect that replaces the potential mediator with a random draw, which is independent of the potential outcome, from the distribution of the mediator among the non-treated. VanderWeele and

Tchetgen Tchetgen (2017) define interventional effects with time-varying exposures and mediators. The estimand is also used in mediation analysis with multiple mediators (VanderWeele and Vansteelandt, 2014; Daniel et al., 2015).

In this paper, we propose a new set of identification assumptions for the natural mediation effects in the presence of treatment-induced confounding without invoking structural equation models, and the identified expression remains to have a causal interpretation even when certain assumptions are violated. In addition to studying the identification of natural direct effect with treatment-induced confounding, we found that, unlike usual expressions, the efficient influence function contains conditional densities that are not variational independent. We consider a reparameterization based on copulas to address the problem of model incompatibility. The corresponding estimator is quadruply robust, that is, consistent under four types of misspecification of nuisance models.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed identification assumptions and the expression of the identified natural direct effect. We explain the connection between our identification results and that of the interventional direct effect. In Section 3, we propose four moment-type estimators and a quadruply robust estimator. In the process, we derive the efficient influence function (hence the semiparametric efficiency bound) of the identified natural direct effect, propose a variation independent parameterization, and prove the quadruple robustness of the estimator. In Section 4, we use numerical simulations to demonstrate the proposed methods. In Section 5, we apply our method to the 2017 Natality data to estimate the effect of prenatal care on preterm birth mediated by preeclampsia with smoking status during pregnancy being a potential treatment-induced confounder. We conclude the paper with some remarks in Section 6. Technical proofs are given in the supplementary materials.

2 Assumptions and Identification

We denote the treatment as *A*, the outcome as *Y*, the mediator as *M*, the set of treatment-induced confounders as *C*, and the set of pre-treatment or baseline covariates as *X*. All variables may be multivariate. Figure 1 demonstrates the causal diagram, showing that (*C*, *M*) are between the causal pathway of $A \to Y$. The set of covariates *X* is omitted for simplicity because it has arrows to all other variables in the causal diagram.

The potential outcome Y_a and the potential mediator M_a are the values the outcome and the mediator would have taken had the treatment been *a*. Such definitions only make sense when there is a well-defined intervention on the treatment *A* so that it can be set to the value *a* (Rubin, 1974; Rubin, 1978; Robins and Greenland, 1992; Pearl, 2001). Similarly, when there are well-defined interventions for *A*, *C*, and *M*, the potential outcome Y_{acm} is the value the outcome would have taken had the treatment been *a*, the treatment-induced confounder been *c*, and the mediator been *m*. We assume the composition (also called recursive substitution) holds such that $Y_a = Y_{aC_aM_a}$, $Y_{am} = Y_{aC_am}$, and $M_a = M_{aC_a}$. Other conditions are needed as preliminaries of identification: The consistency assumption that implies $C_a = C$ and $M_a = M$ when $A = a$; $Y_{acm} = Y$ when $A = a, C = c, M = m$, and the positivity assumption that implies $f(m \mid A, C, X) > 0, f(c \mid A, X) > 0, f(a \mid X) > 0$ with probability 1 for all *m*, *c*, *a*.

When the treatment is binary, the average natural direct effect on a difference scale is defined as $E[Y_{1M_0} - Y_{0M_0}]$. It depicts the expected effect of the treatment when the mediator is fixed at the value it would have taken had there not been any treatment. A set of assumptions sufficient to identify the natural direct effect in the presence of treatment-induced confounders are given as follows: for all *m*, *c*, and *a*,

Assumption 2.1. $\{Y_{am}, C_a, M_a\} \perp\!\!\!\perp A \mid X$.

Assumption 2.2. $Y_{am} \perp\!\!\!\perp M_a \mid A = a, C_a = c, X$

Assumption 2.3. $E[Y_{1C_1m} - Y_{0C_1m} \mid M_0 = m, X] = E[Y_{1C_1m} - Y_{0C_1m} \mid X]$

Assumption 2.4. $E[Y_{0C_1m} - Y_{0C_0m} \mid M_0 = m, X] = E[Y_{0C_1m} - Y_{0C_0m} \mid X]$

Assumptions 2.1 and 2.2 are ignorability assumptions that are implied by the assumptions of no unmeasured confounding between the treatment and post-treatment variables, and between the mediator and the outcome respectively.

Assumptions 2.3 and 2.4 imply that there is no additional heterogeneity in a direct effect of *A*, or in a pure indirect effect of *A* that goes through *C* across levels of M_0. To put it in the data example, where the treatment is adequate prenatal care, the treatment-induced confounder is smoking during pregnancy, the mediator is preeclampsia, and the outcome is preterm birth, the assumption 2.3 implies that the direct effect of adequate prenatal care on preterm birth that goes through neither smoking nor preeclampsia is the same among those who would or would not get preeclampsia without adequate prenatal care. Similar interpretation can be made with Assumption 2.4, which concerns the effect that goes only through smoking during pregnancy.

Theorem 2.1. *Under assumptions 2.1–2.4, the natural direct effect* $E[Y_{1M_0} - Y_{0M_0}]$ *is identified as follows:*

$$\Delta \equiv E_X(E_{M=m|A=0,X}\{E_{C|A=1,X}E[Y \mid A=1, C, M=m, X] - E_{C|A=0,X}E[Y \mid A=1, C, M=m, X]\}). \quad (1)$$

Interestingly, our identification result gives the same empirical expression as the interventional effect (VanderWeele et al., 2014; VanderWeele and Tchetgen Tchetgen, 2017). The interventional direct effect is defined by replacing the potential mediator with a random draw from the distribution of the potential mediator M_0 that is independent of the potential outcomes, and requires only Assumptions 2.1 and 2.2 for identification. In fact, when M_0 is being replaced by

a random draw $M_0^\star$, Assumptions 2.3 and 2.4 are immediately satisfied because $M_0^\star$ is independent of $\{Y_{acm}, C_{a'}\}$.

Note that assumptions 2.1 and 2.2 are "single world" assumptions (Richardson and Robins, 2013) because the treatment is set to be the same in all post-treatment variables in these assumptions. For "single-world" assumptions, one can conceptualize an ideal experiment that intervenes on *A* and *M* such that assumptions 2.1 and 2.2 hold, so as to such an experiment to verify the results from non-experimental data. However, identification of natural direct effect will inevitably involve "cross-world" assumptions where no experiments can be designed to satisfy those assumptions, even when there is no treatment-induced confounders (Imai et al., 2010; Petersen et al., 2006).

Although assumptions 2.3 and 2.4 are "cross-world" assumptions, they do not involve many "cross-world" independence assumptions as in NPSEM-IE (independence between Y_{acm}, $M_{a'c'}$ and $C_{a''}$ for any a, a', a'', c, c' and m), and additional assumptions such as independence between C_0 and C_1 as studied in Robins and Richardson (2010). In addition, a sensitivity analysis of these assumptions are given in Section 6. When there is no treatment-induced confounder, *C* becomes part of *X*, assumption 2.4 becomes redundant, and assumption 2.3 reduces to that in Petersen et al. (2006). The identification of the natural direct effect becomes

$$E_X[E_{M=m|A=0,X}\{E(Y \mid A=1, M=m, X) - E(Y \mid A=0, M=m, X)\}],$$

which is the same empirical expression as the natural direct effect identified by the sequential ignorability assumptions (Pearl, 2001; Imai et al., 2010).

3 Semiparametric Inference

3.1 Moment-type Estimators

Denote the identified expression of the natural direct effect in Theorem 2.1 as Δ, which is the estimand of interest for the remaining sections. The observed

independent samples are $(X_i, A_i, C_i, M_i, Y_i), i = 1, \ldots, n$. With slight abuse of notation, the density (mass) functions are denoted by *f*. The estimand Δ can be represented in four alternative ways, each leading to a possible estimator.

Theorem 3.1. $\Delta = \Delta_1 = \Delta_2 = \Delta_3 = \Delta_4$, *where*

$$\Delta_1 = E_{X,A,C,M,Y}\left\{\frac{2A-1}{f(A \mid X)}\frac{f(M \mid A=0, X)}{f(M \mid A, C, X)}Y\right\},$$
$$\Delta_2 = E_{X,A,C}\left\{\frac{2A-1}{f(A \mid X)}\eta_{C,X}(A)\right\},$$
$$\Delta_3 = E_{X,A,M}\left\{\frac{1-A}{f(A=0 \mid X)}(\gamma_{M,X}(1) - \gamma_{M,X}(0))\right\},$$
$$\Delta_4 = E_X\left\{\tau_X(1) - \tau_X(0)\right\},$$

where

$$\eta_{C,X}(a) = \int E(Y \mid A=a, m, C, X) f(m \mid A=0, X) \mathrm{d}m,$$
$$\tau_X(a) = \int E(Y \mid A=a, m, c, X) f(m \mid A=0, X) f(c \mid A=a, X) \mathrm{d}m \mathrm{d}c,$$
$$\gamma_{M,X}(a) = \int E(Y \mid A=a, M, c, X) f(c \mid A=a, X) dc.$$

The first moment-based estimator Δ_1 is a fully weighted version of the target estimand. The weight is given by the product of the inverse probability weights, and the density ratio between the marginalized density of *M* without treatment and a conditional density of *M*. Intuitively, this density ratio creates a pseudo population in which the distribution of *M* follows $f(M \mid A=0, X)$. The last moment-based estimator Δ_4 is a fully marginalized version of the target estimand that has a similar form as the mediation g-formula (Robins, 1986; Tchetgen Tchetgen and VanderWeele, 2014). Both Δ_2 and Δ_3 are partially marginalized, partially weighted versions of the target estimand. They are both inverse probability weighted marginalized expectations. Compared with the case when treatment-induced confounders are absent,

Tchetgen Tchetgen and Shpitser (2012) proposed a fully weighted estimator, a fully marginalized estimator and one partially marginalized estimator.

Based on different representations of Δ, we consider four estimators $\hat{\Delta}_1, \hat{\Delta}_2, \hat{\Delta}_3$, and $\hat{\Delta}_4$ that replace conditional expectations or densities in $\Delta_1, \Delta_2, \Delta_3$, and Δ_4 with their estimates and the outer expectation by the empirical average. When *Y*, *M*, and *C* are discrete and low dimensional, $\hat{f}(Y \mid A, M, C, X), \hat{f}(M \mid A, C, X), \hat{f}(C \mid A, X)$, and $\hat{f}(A \mid X)$ can be empirical probability mass functions, and $E(Y \mid A, M, C, X)$ is the expectation under $\hat{f}(Y \mid A, M, C, X)$. The integrals in the estimators become finite sums and the four estimators are nonparametric. In practice, however, *M* and *C* are likely to be multivariate and continuous, thus we use parametric models for the purpose of dimension reduction. The four estimators are consistent when nuisance parameters for each part of them are consistently estimated. In particular, with the rest of the models unrestricted, $\hat{\Delta}_1$ is consistent and asymptotically normal when $f(A \mid X), f(M \mid A = 0, X)$ and $f(M \mid A, C, X)$ are correctly specified, $\hat{\Delta}_2$ is consistent and asymptotically normal when $f(A \mid X), E(Y \mid A, M, C, X)$, and $f(M \mid A = 0, X)$ are correctly specified, $\hat{\Delta}_3$ is consistent and asymptotically normal when $f(A \mid X), E(Y \mid A, M, C, X)$, and $f(C \mid A, X)$ are correctly specified, and $\hat{\Delta}_4$ is consistent and asymptotically normal when $E(Y \mid A, M, C, X), f(M \mid A = 0, X)$, and $f(C \mid A, X)$ are correctly specified. When the outcome model is linear and thus collapsible, or when mediators and treatment-induced confounders are categorical, numerical integration is not necessary for the calculation of integrals η, τ, and γ, since the expressions can be simplified. When the outcome model is non-collapsible, *e.g.* a logistic regression model, and the mediator and/or the treatment-induced confounder are continuous, we need to use numerical integration for computation.

3.2 Efficient Influence Function and the Quadruply Robust Estimator

Next, we derive the efficient influence function of Δ under a nonparametric model $\mathcal{M}_{non}$, which does not impose constraints on the observed data.

Theorem 3.2. *The efficient influence function of Δ in $\mathcal{M}_{non}$ is:*

$$S_{\Delta}^{\text{eff}} = \frac{2A-1}{f(A|X)} \frac{f(M|A=0,X)}{f(M|A,C,X)} (Y - E[Y|A,M,C,X]) +$$
$$\frac{2A-1}{f(A|X)} \eta_{C,X}(A) - \frac{2A-1}{f(A|X)} \tau_X(A) + \frac{1-A}{f(A|X)} \{\gamma_{M,X}(1) - \gamma_{M,X}(0)\}$$
$$+\left(1 - \frac{1-A}{f(A|X)}\right)\{\tau_X(1) - \tau_X(0)\} - \Delta.$$

Hence, the semiparametric efficiency bound for the estimation of Δ in $\mathcal{M}_{non}$ is $E[S_{\Delta}^{\text{eff}} S_{\Delta}^{\text{eff}T}]$, and the asymptotic variance of any regular asymptotic linear estimator of Δ in $\mathcal{M}_{non}$ must be greater than or equal to the bound.

The efficient influence function is a function of $f(A|X)$, $f(C|A,X)$, $f(M|A,C,X)$ and $E(Y|A,M,C,X)$. While we may posit parametric working models for these functions, a complication arises because $f(C|A,X)$, $f(M|A,X)$, and $f(M|A,C,X)$ are not variation independent, and therefore model incompatibility may occur. Richardson et al. (2017) point out that the multiple robustness property is relevant only when model incompatibility can be avoided.

We consider reparameterizing the joint distribution $f(M,C|A,X)$ into three parts: the two margins conditioned on *A* and *X*: $f(M|A,X), f(C|A,X)$ and their dependence structure modeled using a copula condition on *A* and *X*. A copula is a multivariate cumulative distribution function with uniformly distributed margins on $[0,1]$. A more detailed discussion on copulas is given by Joe (1997), Nelsen (2007), and Jaworski et al. (2010). For notational simplicity, we consider univariate *M* and *C*, and a bivariate conditional copula with support contained in $[0,1]^2$:

$$\mathcal{C}(u_1,u_2) = P(U_1 \le u_1, U_2 \le u_2),$$

where $P(U_1 \le u_1) = u_1$. Sklar's theorem (Sklar, 1959) allows separate modeling of these three parts. In other words, the joint distribution $F(M,C|A,X)$ is uniquely

determined by $f(M \mid A, X), f(C \mid A, X)$, and $\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ that can be modeled independently. Examples of copulas for continuous, binary, and mixed continuous-binary *M* and *C* are given in the supplementary materials. In multivariate cases, the vine pair copula construction (Panagiotelis et al., 2012) can be used to construct the joint distribution.

Let P_n be the empirical measure. With the variation independent parameterization, we construct a locally efficient estimator based on the following estimating equation:

$$P_n(\hat{S}_{\Delta}^{\text{eff}}(\hat{\Delta}_{quad})) = 0.$$

$\hat{S}_{\Delta}^{\text{eff}}$ is evaluated where all components of the influence function are replaced by their parametric working model: $f(a \mid X)$ is replaced by $f^{\text{par}}(a \mid X), f(c \mid A, X)$ is replaced by $f^{\text{par}}(c \mid A, X), f(m \mid A, X)$ is replaced by $f^{\text{par}}(m \mid A, X)$, and $E(Y \mid A, M, C, X)$ is replaced by $E^{\text{par}}(Y \mid A, M, C, X)$. In particular, $f(m, c \mid A, X)$ is replaced by $f^{\text{par}}(m, c \mid A, X)$, which is modeled by the two marginal distributions $f^{\text{par}}(m \mid A, X), f^{\text{par}}(c \mid A, X)$, and the copula $\mathcal{C}^{\text{par}}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$. Therefore, $\hat{\Delta}_{quad}$ takes the following form:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{2A_i-1}{\hat{f}^{\text{par}}(A_i \mid X_i)}\frac{\hat{f}^{\text{par}}(M_i \mid A_i=0, X_i)\hat{f}^{\text{par}}(C_i \mid A_i, X_i)}{\hat{f}^{\text{par}}(M_i, C_i \mid A_i, X_i)}\Big\{Y_i - E^{\text{par}}(Y_i \mid A_i, M_i, C_i, X_i)\Big\} \\
&+\frac{2A_i-1}{\hat{f}^{\text{par}}(A_i \mid X_i)}\Big\{\hat{\eta}^{\text{par}}_{M_i,X_i}(A_i) - \hat{\tau}^{\text{par}}_{X_i}(A_i)\Big\} \\
&+\frac{1-A_i}{\hat{f}^{\text{par}}(A_i=0 \mid X_i)}\Big[\Big\{\hat{\gamma}^{\text{par}}_{M_i,X_i}(1) - \hat{\gamma}^{\text{par}}_{M_i,X_i}(0)\Big\} - \Big\{\hat{\tau}^{\text{par}}_{X_i}(1) - \hat{\tau}^{\text{par}}_{X_i}(0)\Big\}\Big] \\
&+\Big\{\hat{\tau}^{\text{par}}_{X_i}(1) - \hat{\tau}^{\text{par}}_{X_i}(0)\Big\}\Big).
\end{aligned} \tag{2}$$

This estimator is quadruply robust in the sense that only one out of four sets of models needs to be correctly specified for it to be consistent and asymptotically normal as given in Theorem 3.3.

Theorem 3.3. *The estimator* $\hat{\Delta}_{quad}$ *is consistent and asymptotically normal under some mild regularity conditions discussed in the supplementary materials if one of the following four conditions holds:*

1. $\mathcal{M}_1 : f^{\text{par}}(A \mid X), f^{\text{par}}(C \mid A, X), f^{\text{par}}(M \mid A, X), \mathcal{C}^{\text{par}}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ *are correctly specified.*
2. $\mathcal{M}_2 : f^{\text{par}}(A \mid X), f^{\text{par}}(M \mid A, X), E^{\text{par}}(Y \mid A, M, C, X)$ *are correctly specified.*
3. $\mathcal{M}_3 : f^{\text{par}}(A \mid X), f^{\text{par}}(C \mid A, X), E^{\text{par}}(Y \mid A, M, C, X)$ *are correctly specified.*
4. $\mathcal{M}_4$ *:* $f^{\text{par}}(M \mid A, X), f^{\text{par}}(C \mid A, X), \mathcal{C}^{\text{par}}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X), E^{\text{par}}(Y \mid A, M, C, X)$ *are correctly specified.*

It is locally semiparametric efficient in the sense that it achieves the semiparametric efficiency bound at the intersection of the submodels where all four conditions hold, that is, at $\mathcal{M}_{intersection} = \mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3 \cap \mathcal{M}_4$.

Due to the complexity of the quadruply robust estimator, the multiple robustness is not easily seen directly from its form. We explicitly illustrate the robustness of the estimator under $\mathcal{M}_1$ as an example. The large sample limit of the estimating equation for the quadruply robust estimator $\hat{\Delta}_{quad}$ can be written as the sum of four parts:

1. $E\left\{\dfrac{2A-1}{f_{A|X}(A \mid X)} \dfrac{f_{M|A,X}(M \mid A=0, X)}{f_{M|A,C,X}(M \mid A, C, X)} Y\right\} - \Delta$,
2. $E\left\{\dfrac{2A-1}{f_{A|X}(A \mid X)} \eta^*_{C,X}(A)\right\} - E\left\{\dfrac{2A-1}{f_{A|X}(A \mid X)} \dfrac{f_{M|A,X}(M \mid A=0, X)}{f_{M|A,C,X}(M \mid A, C, X)} E^*(Y \mid A, M, C, X)\right\}$,
3. $E\{\tau^*_X(1) - \tau^*_X(0)\} - E\left\{\dfrac{2A-1}{f_{A|X}(A \mid X)} \tau^*_X(A)\right\}$,
4. $E\left(\dfrac{\mathbf{1}(A=0)}{f_{A|X}(A=0 \mid X)}[\{\gamma^*_{M,X}(1) - \gamma^*_{M,X}(0)\} - \{\tau^*_X(1) - \tau^*_X(0)\}]\right)$,

where quantities with superscript * represents the components that are incorrectly specified under $\mathcal{M}_1$. Note that the except for the first term, all other terms include misspecified quantities. We proved that each of the four parts equals to 0 in the supplementary materials. Similarly, under each of $\mathcal{M}_2, \mathcal{M}_3$ and $\mathcal{M}_4$, the large sample limit of the estimating equation can be rewritten into sum of four parts where one of it contains the correctly specified quantities only and the other three contain mis-specified quantities, but that all four parts can be shown to be 0. Details are given in the supplementary materials. Notice that the estimators proposed in section 3.1 are such that $\hat{\Delta}_1$, whose estimation can be conducted using the copula parameterization, is only consistent under $\mathcal{M}_1$, Δ_2 is only consistent under $\mathcal{M}_2$, $\hat{\Delta}_3$ is only consistent under $\mathcal{M}_3$, and $\hat{\Delta}_4$ is only consistent under $\mathcal{M}_4$. In contrast, the quadruply robust estimator $\hat{\Delta}_{quad}$ remains consistent under four types of misspecification, which offers more modeling flexibility. In other words, $\hat{\Delta}_{quad}$ is consistent and asymptotically normal at the intersection submodel.

Remark 1. *Since the quadruply robust estimator involves weighting and the weights could be unbounded when models are mis-specified, the resulting estimator can be unstable when none of $\mathcal{M}_1$ to $\mathcal{M}_4$ holds (Kang et al., 2007). To improve the stability of the weights thereby improve the finite sample performance of the quadruply robust estimator, we extend the methods proposed in Robins et al. (2007) and Tchetgen Tchetgen and Shpitser (2012) to our setting to construct a stabilized quadruply robust estimator with a certain boundedness property by modifying the estimation procedure of the parametric working models such that the first three terms of (2) are exactly zero. This requires careful construction of estimating equations for working models and needs to be considered case by case. In Section 4.2, we give an exact procedure under a certain simulation setting.*

Remark 2. *While there are other parametrizations of joint densities of (C, M), such as Chen (2007) whose decomposition depends on two conditional*

distributions and a odds ratio function, it appears that these characterizations would not ensure multiple robustness because marginals of C and M are part of the robustness conditions in Theorem 3.3.

3.3 Related estimands

One favorable feature of the natural direct effect is that the average treatment effect, defined as $E[Y_1 - Y_0]$, can be decomposed into the sum of the average natural direct effect and the natural indirect effect: $E[Y_{1M_1} - Y_{1M_0}]$. While the natural indirect effect is not the focus of this paper, similar results can be applied to it since the average treatment effect is identified under assumption 2.1. The natural indirect effect is then identified as the difference between the identified average treatment effect and the natural direct effect identified in Theorem 2.1. We should note, however, that the identified natural indirect effect is different from the interventional indirect effect. This is consistent with the fact that the interventional direct effect and indirect effect do not sum up to be the average treatment effect (Vansteelandt and Daniel, 2017). The semiparametric estimation theory can also be extended for the natural indirect effect. Specifically, we can construct a quadruply robust estimator for the natural indirect effect by the difference between the augmented inverse probability weighted estimator for the average treatment effect (Robins et al., 1994; Robins, 2000, Tsiatis, 2007), and our proposed quadruply robust estimator. The augmented inverse propensity weighted estimator is consistent if either the model for the propensity score or the regression model for the mean outcome is correct. Notice that for each of $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$, the condition for the doubly robust average treatment effect estimator being consistent is satisfied. Therefore the quadruple robustness extends to the natural indirect effect.

Although we studied the natural direct effect defined as a difference in expectations, this effect can also be defined on other scales, such as a ratio scale $E(Y_{1M_0}) / E(Y_{0M_0})$. Because $E(Y_{0M_0}) = E(Y_0)$ is identified under assumption 2.1 and can be estimated by the augmented inverse probability weighted

estimator $\hat{E}(Y_0), E(Y_{1M_0})$ can be estimated by $\hat{\Delta} + \hat{E}(Y_0)$. Therefore, any functions of $E(Y_{1M_0})$ and $E(Y_0)$ can be estimated and the asymptotic variance can be derived using the delta method. However, since the identification assumptions of Δ are given on the difference scale, extra care is needed when interpreting the natural direct and indirect effect defined on other scales.

4 Simulation study

4.1 Demonstration of Theoretical Results

We use numerical simulations to demonstrate the theoretical results derived in the previous section. We compare the finite sample performance of the moment-based estimators given in section 3.1 to the proposed quadruply robust estimator. We generate 1000 samples, each with 1500 independent observations, for both continuous and binary treatment-induced confounder and mediator. We consider the moment estimators $\hat{\Delta}_1, \hat{\Delta}_2, \hat{\Delta}_3, \hat{\Delta}_4$ and the quadruply robust estimator $\hat{\Delta}_{quad}$. Let *expit* denote the function $expit(x) = exp(x)/(1+exp(x))$. The data are generated as follows:

Continuous C and M:

$X \sim N(0,1); P(A=1 \mid X) = expit(-0.4 + 0.6X);$

$\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ is a Gaussian Copula with correlation 0.2,

where $F_{M|A,X}(m) = \Phi(\frac{m-\mu_m}{\sigma_m}), \mu_m = 3 + 2A + 4X, \sigma_m = 5,$

$F_{C|A,X}(c) = \Phi(\frac{c-\mu_c}{\sigma_c}), \mu_c = 1 + 2A + 2X, \sigma_c = 4,$

$Y \sim 1 + 2A + 2M + 3C + 5X + 4AC + 2AM + N(0, 4^2).$

Binary C and M:

$X \sim N(0,1); P(A=1 \mid X) = expit(-0.2+0.3X);$
$\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ is a Plackett Copula with Odds-Ratio $exp(1-2A+3X)$;
where $F_{M|A,X}(m) = p_m^m(1-p_m)^{1-m}, p_m = expit(-0.3-0.2A+0.5X),$
$F_{C|A,X}(c) = p_c^c(1-p_c)^{1-c}, p_c = expit(-0.2-0.1A+0.3X),$
$Y \sim 1+3A+6M+3C+6X+4AC+2AM+N(0,4^2).$

We compare the five estimators under a series of model misspecifications by replacing the baseline covariates X with an independent normally distributed continuous variable X_2 with mean 0 and variance 1. The true natural direct effects are 26.01 and 5.50 for the continuous and binary cases, respectively. Table 1 shows that the simulation results are consistent with the theoretical results derived in the previous sections: when the entire likelihood is correctly specified, all five estimators are consistent; when the conditional expectation of Y is mis-specified, only Δ_1 and Δ_{quad} are consistent; when the parametric model for $f(C \mid A, X)$ is mis-specified, only Δ_2 and Δ_{quad} are consistent; when the parametric model for $f(M \mid A, X)$ is mis-specified, only Δ_3 and Δ_{quad} are consistent; when the propensity score $f(A=1 \mid X)$ is mis-specified, only Δ_4 and Δ_{quad} are consistent. The loss in efficiency for the quadruply robust estimator is relatively small compared to other estimators in all cases. Since Δ_1 consists of a density ratio, it is more variable when the mediator M is continuous, which makes it less preferred even when $\mathcal{M}_1$ is correct. We only present one scenario here, but we ran simulations under different settings and they all gave similar results.

We also include a comparison with the triply robust estimator proposed in Tchetgen Tchetgen and Shpitser (2012), which assumes the absence of treatment-induced confounding. We consider two cases, with bias and standard error multiplied by 100 as in Table 1. First, where C is ignored in the estimation, and only use (Y, M, A, X) in the estimation. The sampling bias (sampling standard error) is $-2577.83\ (164.60)$ for continuous mediators and $-93.10\ (65.84)$ for binary mediators. Next, we erroneously treat C as a pre-treatment covariate and condition on (C, X) in all working models of the triply robust estimator. The

sampling bias (sampling standard error) is $-2192.33\ (172.94)$ for continuous mediators and $-96.49\ (65.60)$ for binary mediators. Therefore, ignoring treatment-induced confounders or treating them as pre-treatment covariates could lead to substantially biased results.

4.2 Practical Violation of Positivity Assumptions

Next we consider a scenario similar to Kang et al. (2007) in which the positivity of the treatment assignment probability is practically violated under model misspecification. The data are simulated as follows:

$X \equiv (Z_1, Z_2, Z_3, Z_4) \sim N(0, I_4),$ where I where

$A \sim Bernoulli(p_a),$ where $p_a = expit(-Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4);$

$C \sim Bernoulli(p_c),\ (where)\, p_c = expit(-1.6 + 2A + Z_1 - 0.8Z_2 + 0.6Z_3 - Z_4);$

$M \sim Bernoulli(p_m),$ where $p_m = expit(-1.5 + 2A + Z_1 - 0.5Z_2 + 0.9Z_3 - Z_4);$

$\mathcal{C}(F_{M|A,X}(m), F_{C|A,X}(c) \mid A, X)$ is a Placket Copula with odds ratio exp(1.2);

$Y \sim 210 + A + M - 50C + 27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4 + N(0, 30^2).$

Instead of observing the $Z's$, we observe transformations of them:

$X_1 = exp(Z_1/2), X_2 = Z_2/\{1 + exp(Z_1)\} + 10, X_3 = (Z_1 Z_3/25 + 0.6)^3, X_4 = (Z_2 + Z_4 + 20)^2.$

Correctly specified model should include the true "latent" covariates (Z_1, Z_2, Z_3, Z_4). Instead we replace them using the "observed" covariates (X_1, X_2, X_3, X_4) in misspecified models. The true natural direct effect is –14.57 in this case.

When models are mis-specified, the weights in the quadruply robust estimator can be unbounded, and the resulting estimator can be unstable when none of $\mathcal{M}_1$ to $\mathcal{M}_4$ holds. In fact, the weights can be unstable even when the models are

well-specified under practical violation of positivity (Westreich and Cole, 2010). As mentioned in Section 3.2, a stabilized quadraply robust estimator would be desired to handle such a case. It can be written in the form $\hat{\Delta}^{\dagger}_{quad} = P_n\{\hat{\tau}^{\dagger}_X(1) - \hat{\tau}^{\dagger}_X(0)\}$, where $\hat{\tau}^{\dagger}_X(1)$ and $\hat{\tau}^{\dagger}_X(1)$ are estimated in a manner that ensures quadruple robustness. Note that the definition of the target estimand is $\Delta = E[\tau_X(1) - \tau_X(0)]$, hence the quadruply robust estimator $\hat{\Delta}^{\dagger}_{quad}$ lies in the range of the target estimand Δ as long as $\hat{\tau}^{\dagger}_X(1) - \hat{\tau}^{\dagger}_X(0)$ lies in the same range. For a continuous *Y* and binary *A*, *C* and *M*, a stabilized estimator is constructed using the following procedure:

1. To estimate $\hat{f}^{par\dagger}(C \mid A = a, X)$ using a weighted logistic regression in the group with treatment *a* with weights $f^{par}(a \mid X)^{-1}$.
2. To estimate $\hat{f}^{par\dagger}(M \mid A = a, X)$ using weighted logistic regression in the group with treatment *a* with weights $f^{par}(a \mid X)^{-1}$.
3. To estimate $E[Y \mid X, M, C, A = a]$ in the group with treatment *a* using weighted least square regression with weights

$$f^{par}(a \mid X)^{-1} \frac{\hat{f}^{par\dagger}(M \mid A = 0, X)\hat{f}^{par\dagger}(C \mid A = a, X)}{\hat{f}^{par\dagger}(M, C \mid A = a, X)},$$

where $\hat{f}^{par\dagger}(M, C \mid A = a, X)$ is estimated using the estimated copula and two marginal distributions $\hat{f}^{par\dagger}(C \mid A = a, X)$ and $\hat{f}^{par\dagger}(M \mid A = a, X)$.

Tables 2 and 3 summarize the simulation results for the four moment based estimators and the quadruply robust estimator for sample sizes 500 and 1000, with 1000 independent replications in each scenario. Note that when two or more models are incorrect, none of $\mathcal{M}_1$ to $\mathcal{M}_4$ holds. In most of these scenarios, the stabilized version of the quadruply robust estimator has smaller bias than all other estimators, including the original quadruply robust estimator. When only model $f(A \mid X)$ is incorrect, $\mathcal{M}_4$ holds so the unstabilized quadraply robust estimator has small estimation bias, but the stabilized estimator reduces the sampling standard errors substantially. When $\mathcal{M}_1$ to $\mathcal{M}_3$ holds and under the

intersection model, the sampling bias and standard derivations of the unstabilized and stabilized quadraply robust estimators are very similar. The stabilized version of the quadruply robust estimator is recommended in practice because it has a better performance both in terms of bias and standard error than the unstabilized version.

5 Data Example

We use the 2017 Natality data (https://wonder.cdc.gov/natality.html) for births occurring within the United States to U.S. residents to illustrate our method. We focus our analysis on the subset of participants that are AIAN (American Indians or Alaskan Native). Subjects with missing data ($< 9.5\%$ of the sample) are excluded. The total number of observations is 27,138.

We are interested in estimating the direct effect of prenatal care (A) on preterm birth (Y) not through preeclampsia (M). As pointed out in the introduction, smoking status during pregnancy (C) is a potential treatment-induced confounder. The adequacy of prenatal care is determined by the Adequacy of Prenatal Care Utilization Index (Kotelchuck, 1994), which depends on the month prenatal care began, the total number of prenatal visits, and the gestational age at the time of delivery. In the AIAN sample, the level of prenatal care is either inadequate or intermediate. Preterm birth is defined using the Obstetric Estimate (OE) (Martin et al., 2015) of the gestational age. The baseline covariates (X) that are potential confounders include maternal demographics: age, education level, and marital status. Assumption 2.3 implies that the direct effect of prenatal care on preterm birth (that goes through neither smoking nor preelampsia) is the same among those who would get preelampsia without adequate prenatal care, and those who would not. Similarly, Assumption 2.4 implies that the mediated effect of prenatal care through smoking is the same among those who would get preelampsia without adequate prenatal care, and those who would not. If these two assumptions are violated, meaning that the potential preelampsia status without adequate prenatal care modifies either the direct effect of prenatal care

or its mediated effect through smoking, then the estimated effects can be interpreted as interventional effects, as explained in section 2.

Since both the smoking status and the preeclampsia status are binary, we use the Plackett copula with a cross-ratio (odds ratio) specified using a log link. Logistic regression models are used for the binary treatment and outcome, as well as the distributions of C and M given A and X. The parameters of the copula are estimated by the maximum likelihood method. The bootstrap confidence intervals are computed for the purpose of inference.

The estimated direct effect of better prenatal care (intermediate care versus inadequate care) not through preeclampsia decreases the risk of preterm birth by 2.5% (1.6%, 3.4%), leaving a tiny indirect effect through preeclampsia that increases the risk of preterm birth by 0.15% (0.07%, 0.23%). The moment-type estimators give similar results (Table 4). This is consistent with VanderWeele et al. (2014) who studied this problem on a different population.

6 Discussion

In this paper, we identify the natural direct effect in the presence of treatment-induced confounding, and derive semiparametric bounds and propose a quadruply robust estimator. Our method can be applied to continuous, categorical, and multivariate outcomes, and to mediators and treatment-induced confounders.

When identification assumptions may be violated, sensitivity analysis can be useful to assess how vulnerable the estimator would be. Inspired by Vansteelandt and VanderWeele (2012) and VanderWeele and Chiba (2014), we can consider the following two sensitivity functions:

$$q_m(M_0, X) = E[Y_{1C_1m} - Y_{0C_1m} \mid M_0 = m, X] - E[Y_{1C_1m} - Y_{0C_1m} \mid M_0, X],$$
$$l_m(M_0, X) = E[Y_{0C_1m} - Y_{0C_0m} \mid M_0 = m, X] - E[Y_{0C_1m} - Y_{0C_0m} \mid M_0, X].$$

The former captures the heterogeneity in the direct effect of the treatment across different subgroups defined by potential mediators under control and baseline confounders, and the latter captures the heterogeneity in the indirect effect of the treatment through the treatment-induced confounder across different subgroups. Let $Q_m(X) = E(q_m(M_0, X) \mid X) = E(q_m(M, X) \mid A = 0, X)$ and $L_m(X) = E(l_m(M_0, X) \mid X) = E(l_m(M, X) \mid A = 0, X)$, then assumuption 2.3 and 2.4 corresponds to $Q_m = 0$ and $L_m = 0$ respectively. Assuming the sensitivity functions to be known, the natural direct effect can be identified as $\Delta + E(Q_m(X)) + E(L_m(X))$.

Another possible direction is to develop bounds when identification assumptions are relaxed. However, bounds are often developed for categorical data where linear inequality constraints may be specified. Bounds on the natural direct effect for a binary mediator are given by Robins and Richardson (2010), which are extended by Tchetgen and Phiri (2014) in the presence of treatment-induced confounding, and are further extended to the polytomous mediator by Miles et al. (2015).

As Robins and Richardson (2010) point out, different assumptions give different identifying expressions. It is sometimes not clear how scientists can choose an identification assumption when it lacks scientific justification, because they are often not refutable even by experiments. Our identified expression has the advantage that even when the no additional effect heterogeneity assumptions are inappropriate, it can still be interpreted as the interventional effect, to which the semiparametric theory and the quadruply robust estimator are still applicable.

Finally, a reviewer inquire us about the possibility of handling high-dimensional (M, C) using this method. In such cases, the proposed semiparametric framework will require modeling of the marginal distributions of M and C and the joint distribution through a copula function, which could be a difficult task with high dimensional M and C. Even under a linear structural equation model with

independent errors, bootstrapping may have non-standard performance as in El Karoui and Purdom (2018), and would require future theoretical investigation.

Supplementary Materials

A written supplementary material contains the proof of Theorem 2.1, 3.1, 3.2 and 3.3, examples of copula for discrete and continuous data, further discussions of the stabilization procedure given in Section 4.2, and additional simulations. Supplementary files containing the codes for numerical results are also provided.

Acknowledgement

The authors thank an associate editor and two anonymous reviewers for their insightful suggestions. This research was partially supported by the US National Institutes of Health grants R01 HL 122212, U01 AG 016976, U24 AG 072122, the US National Science Foundation grant DMS 17 11952.

References

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of International Joint Conference on Artificial Intelligence*.

Chen, H. Y. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2):413–421.

Daniel, R., De Stavola, B., Cousens, S., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14.

El Karoui, N. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, 19(1):170–235.

Hafeman, D. M. and VanderWeele, T. J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology*, pages 753–764.

Imai, K., Keele, L., Yamamoto, T., et al. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 25(1):51–71.

Imai, K. and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171.

Jaworski, P., Durante, F., Hardle, W. K., and Rychlik, T. (2010). *Copula theory and its applications*, volume 198. Springer.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Kotelchuck, M. (1994). An evaluation of the Kessner adequacy of prenatal care index and a proposed adequacy of prenatal care utilization index. *American journal of public health*, 84(9):1414–1420.

Martin, J. A., Osterman, M., Kirmeyer, S., and Gregory, E. (2015). Measuring gestational age in vital statistics data: transitioning to the obstetric estimate. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 64(5):1–20.

Miles, C. H., Kanki, P., Meloni, S., and Tchetgen, E. J. T. (2015). On partial identification of the pure direct effect. *arXiv preprint arXiv:1509.01652*.

Miles, C. H., Shpitser, I., Kanki, P., Meloni, S., and Tchetgen Tchetgen, E. J. (2020). On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, 107(1):159–172.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Panagiotelis, A., Czado, C., and Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072.

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

Pearl, J. (2009). *Causality*. Cambridge university press.

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

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128.

Richardson, T. S., Robins, J. M., and Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, 112(519):1121–1130.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559.

Robins, J. M. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. *Computation, causation, and discovery*, pages 349–405.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. *Oxford Statistical Science Series*, pages 70–82.

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

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035.

Sklar, A. (1959). Fonctions de reprtition an dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris*, 8:229–231.

Tchetgen, E. J. T. and Phiri, K. (2014). Bounds for pure direct effect. *Epidemiology (Cambridge, Mass.)*, 25(5):775–776.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40(3):1816–1845.

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. (2014). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology (Cambridge, Mass.)*, 25(2):282–291.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

VanderWeele, T. J. and Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, biostatistics, and public health*, 11(2):e9027.

VanderWeele, T. J. and Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B*, 79(3):917–938.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2(4):457–468.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306.

Vansteelandt, S. and Daniel, R. M. (2017). Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258–265.

Vansteelandt, S. and VanderWeele, T. J. (2012). Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics*, 68(4):1019–1027.

Westreich, D. and Cole, S. R. (2010). Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171(6):674–677.

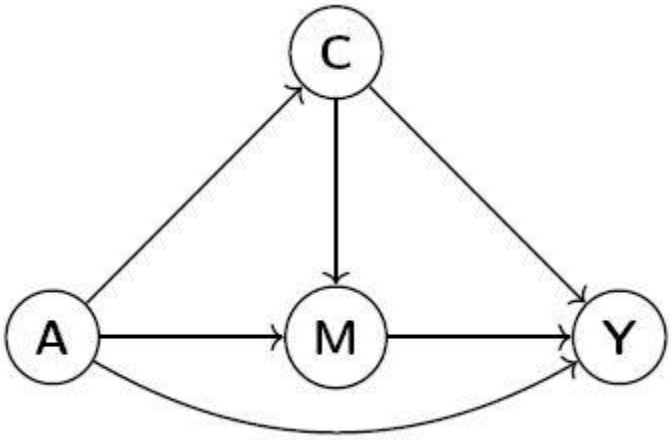

Fig. 1 The Causal Diagram with Treatment-induced Confounding

Table 1 Simulation Results: $100 \times$ Bias ($100 \times$ Standard Error)

Continuous *C*, *M*					
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$
All correct	-5 (191)	-3 (151)	-4 (141)	-3 (139)	-4 (144)
$\mathcal{M}_1$ is correct	10 (180)	87 (141)	89 (137)	88 (135)	2 (138)
$\mathcal{M}_2$ is correct	77 (176)	3 (149)	599 (141)	600 (140)	4 (144)
$\mathcal{M}_3$ is correct	-1390 (369)	-189 (135)	-6 (134)	-187 (133)	-6 (135)
$\mathcal{M}_4$ is correct	1589 (220)	1587 (187)	-359 (143)	4 (134)	4 (134)
Binary *C*, *M*					
	$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$
All correct	1 (154)	-1 (26)	-1 (26)	-1 (26)	1 (49)
$\mathcal{M}_1$ is correct	-2 (115)	176 (40)	175 (40)	176 (40)	3 (115)
$\mathcal{M}_2$ is correct	44 (27)	1 (26)	1 (25)	12 (26)	1(26)
$\mathcal{M}_3$ is correct	22 (30)	-4 (24)	-1 (24)	-4 (24)	-1 (25)
$\mathcal{M}_4$ is correct	245 (44)	210 (46)	-9 (27)	-2 (27)	-2 (27)

Table 2 Simulation with sample size 500. Bias (Standard Error).

Estimator		$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$	$\hat{\Delta}^{\dagger}_{quad}$
all correct	bias	0.04	-0.05	0.11	0.12	-0.05	-0.03
	s.e.	24.17	20.73	3.80	3.75	4.64	4.41
incorrect *A*	bias	34.06	37.93	-0.16	0.23	-0.62	0.27
	s.e.	145.11	174.94	3.67	3.66	18.35	5.29
incorrect *C*	bias	3.22	-0.49	1.88	1.88	0.18	0.18
	s.e.	25.53	21.40	3.80	3.78	4.67	4.43
incorrect *M*	bias	-9.29	-0.32	0.21	0.21	0.07	0.08
	s.e.	22.19	21.13	3.73	3.67	4.14	4.41
incorrect *Y*	bias	-0.14	-8.67	-9.31	-9.37	-0.33	0.52
	s.e.	23.53	21.41	4.21	3.99	4.83	4.75
incorrect *A*, *Y*	bias	98.40	108.50	-9.31	-9.08	-23.20	-4.58
	s.e.	1660.12	2067.75	4.00	3.99	374.43	5.42
incorrect *A*, *C*	bias	32.82	37.79	1.62	1.66	1.08	1.09
	s.e.	121.95	136.16	3.68	3.69	13.66	5.32
incorrect *C*, *Y*	bias	-4.55	-8.74	-7.73	-7.76	0.89	0.50
	s.e.	25.56	21.42	4.04	3.84	5.16	5.05
incorrect *M*, *A*	bias	27.58	33.61	-0.59	-0.19	-1.52	-0.27
	s.e.	280.53	318.43	3.70	3.68	40.94	5.95
incorrect *M*, *C*	bias	-7.26	0.70	1.75	1.72	0.28	0.26
	s.e.	22.45	20.79	3.89	3.82	4.40	4.53
incorrect *M*, *Y*	bias	-9.96	-9.53	-9.26	-9.26	-3.95	-4.21
	s.e.	21.44	20.71	4.31	4.06	4.54	4.74
incorrect *M*, *A*, *C*	bias	27.55	32.56	1.61	1.64	1.14	1.22

Estimator		$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$	$\hat{\Delta}^{\dagger}_{quad}$
	s.e.	124.22	129.27	3.66	3.66	20.28	5.41
incorrect *A*,*C*,*Y*	bias	38.36	38.20	-7.86	-7.95	-7.27	-2.86
	s.e.	194.37	283.71	3.85	3.88	68.42	5.57
incorrect *M*,*C*,*Y*	bias	-8.89	-10.00	-7.93	-7.89	-3.44	-3.36
	s.e.	23.38	22.39	4.23	3.90	5.04	5.05
incorrect *M*,*A*,*Y*	bias	32.68	34.95	-9.54	-9.33	-14.88	-8.38
	s.e.	241.50	183.98	3.71	3.71	71.32	5.23
all incorrect	bias	21.59	24.25	-7.58	-7.91	-10.50	-6.40
	s.e.	88.88	97.63	3.79	3.85	23.99	5.17

Table 3 Simulation with sample size 1000. Bias (Standard Error).

Estimator		$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$	$\hat{\Delta}^{\dagger}_{quad}$
all correct	bias	0.02	0.02	0.10	0.10	0.04	0.05
	s.e.	17.00	14.87	2.64	2.62	3.20	3.13
incorrect *A*	bias	33.31	35.45	-0.36	0.03	0.29	0.01
	s.e.	121.68	127.79	2.55	2.54	7.97	4.31
incorrect *C*	bias	3.54	-0.03	1.64	1.64	0.004	0.02
	s.e.	16.92	14.36	2.64	2.60	3.36	3.19
incorrect *M*	bias	-8.77	0.73	0.26	0.25	0.16	0.16
	s.e.	15.46	14.68	2.75	2.68	3.06	3.14
incorrect *Y*	bias	0.95	-8.45	-9.25	-9.32	0.07	0.07
	s.e.	16.32	14.32	2.90	2.80	3.49	3.42
incorrect *A*, *Y*	bias	60.01	75.18	-0.35	-9.14	-18.24	-4.00
	s.e.	501.58	771.00	2.91	2.89	182.44	4.53
incorrect *A*, *C*	bias	32.99	31.36	1.65	1.69	1.84	1.52
	s.e.	66.38	64.62	2.59	2.59	7.17	4.38
incorrect *C*, *Y*	bias	-1.10	-8.94	-7.80	-7.83	0.86	0.45
	s.e.	17.12	14.71	2.86	2.75	3.54	3.45
incorrect *M*, *A*	bias	38.63	44.16	-0.26	0.14	-0.89	0.03
	s.e.	202.86	224.69	2.56	2.56	30.09	4.75
incorrect *M*, *C*	bias	-8.28	0.15	1.49	1.49	-0.14	-0.16
	s.e.	16.15	15.43	2.64	2.58	2.91	2.98
incorrect *M*, *Y*	bias	-8.57	-8.33	-9.13	-9.16	-3.92	-4.03
	s.e.	16.02	15.64	2.93	2.78	3.28	3.45
incorrect *M*,*A*,*C*	bias	61.96	70.48	1.69	1.73	-1.42	1.37

Estimator		$\hat{\Delta}_1$	$\hat{\Delta}_2$	$\hat{\Delta}_3$	$\hat{\Delta}_4$	$\hat{\Delta}_{quad}$	$\hat{\Delta}^{\dagger}_{quad}$
	s.e.	670.96	748.10	2.65	2.65	117.83	4.64
incorrect *A*,*C*,*Y*	bias	33.15	32.34	-7.78	-7.88	-6.13	-2.46
	s.e.	60.71	81.40	2.64	2.66	14.31	4.16
incorrect *M*,*C*,*Y*	bias	-7.78	-8.88	-7.88	-7.90	-3.40	-3.30
	s.e.	15.72	15.29	2.79	2.66	3.23	3.44
incorrect *M*,*A*,*Y*	bias	58.51	74.34	-9.31	-9.11	-25.19	-7.81
	s.e.	560.11	822.28	2.82	2.83	264.87	4.66
all incorrect	bias	55.19	74.10	-7.41	-7.75	-26.31	-5.80
	s.e.	744.14	1132.02	2.63	2.68	388.65	4.63

Table 4 Estimation of Direct Effect of Better Prenatal Care on Preterm Birth

Estimator	Direct Effect Estimate	Bootstrap 95% CI
$\hat{\Delta}_1$	0.026	(0.016, 0.036)
$\hat{\Delta}_2$	0.028	(0.018, 0.037)
$\hat{\Delta}_3$	0.027	(0.018, 0.036)
$\hat{\Delta}_4$	0.027	(0.018, 0.036)
$\hat{\Delta}_{quad}$	0.025	(0.016, 0.034)