Multiply Robust Causal Inference with Double Negative Control Adjustment for Categorical Unmeasured Confounding

Xu Shi[1], Wang Miao[2], Jennifer C. Nelson[3], and Eric J. Tchetgen Tchetgen[4]

[1]Department of Biostatistics, University of Michigan
[2]Guanghua School of Management, Peking University
[3]Kaiser Permanente Washington Health Research Institute
[4]Department of Statistics, the Wharton School, University of Pennsylvania

Abstract

Unmeasured confounding is a threat to causal inference in observational studies. In recent years, use of negative controls to mitigate unmeasured confounding has gained increasing recognition and popularity. Negative controls have a longstanding tradition in laboratory sciences and epidemiology to rule out non-causal explanations, although they have been used primarily for bias detection. Recently, Miao et al. (2018) have described sufficient conditions under which a pair of negative control exposure and outcome variables can be used to nonparametrically identify the average treatment effect (ATE) from observational data subject to uncontrolled confounding. In this paper, we establish nonparametric identification of the ATE under weaker conditions in the case of categorical unmeasured confounding and negative control variables. We also provide a general semiparametric framework for obtaining inferences about the ATE while leveraging information about a possibly large number of measured covariates. In particular, we derive the semiparametric efficiency bound in the nonparametric model, and we propose multiply robust and locally efficient estimators when nonparametric estimation may not be feasible. We assess the finite sample performance of our methods in extensive simulation studies. Finally, we illustrate our methods with an application to the postlicensure surveillance of vaccine safety among children.

Keywords: causal inference, negative control, semiparametric inference, unmeasured confounding.

1 Introduction

Causal inference in observational studies often relies on the assumption of no unmeasured confounding. However, as often the case in practice, when this assumption is violated, uncontrolled confounding can lead to biased estimates and invalid conclusions. Various methods have been proposed to detect and control for unmeasured confounding, among which use of negative controls has recently gained increasing recognition and popularity. Negative controls have a longstanding tradition in laboratory sciences and epidemiology to rule out non-causal explanation of empirical findings (Rosenbaum, 1989; Weiss, 2002; Lipsitch et al., 2010; Glass, 2014). Specifically, a negative control outcome is an outcome known not to be causally affected by the treatment of interest. Likewise, a negative control exposure is an exposure that does not causally affect the outcome of interest. To the extent possible, both negative control exposure and outcome variables should be selected such that they share a common confounding mechanism as the exposure and outcome variables of primary interest. For example, in a study about the effect of influenza vaccination on influenza hospitalization, injury/trauma hospitalization was considered as a negative control outcome as it is not causally affected by influenza vaccination, but may be subject to the same confounding mechanism mainly driven by health-seeking behavior (Jackson et al., 2005). In this case, a non-null effect of the influenza vaccination against the negative control outcome amounts to compelling evidence of potential bias due to uncontrolled confounding. Another prominent example is the use of paternal exposure as a negative control exposure when determining the effect of maternal exposure during pregnancy on offspring health outcomes. Paternal exposure may have a similar association with the outcome as that of maternal exposure if there is hidden genetic or household-level confounding (Davey Smith, 2008, 2012; Lipsitch et al., 2012).

There is a growing literature on use of negative controls to mitigate confounding bias. Rosenbaum (1992) considered testing and sensitivity analysis for unmeasured confounding by comparing matched treatment and control groups with respect to an unaffected outcome. Tchetgen Tchetgen (2013) developed an outcome calibration approach based on the idea that the counterfactual primary outcomes can stand as a proxy for unmeasured confounders and suffice to account for confounding of the exposure–negative control outcome association. Schuemie et al. (2014) proposed a p-value calibration approach by deriving an empirical null distribution of treatment effect using a collection of negative controls. Sofer et al. (2016) generalized the difference-in-difference approach to the broader context of negative control outcome by allowing different scales for primary and negative control outcomes under a monotonicity assumption. In genetic studies, Gagnon-Bartsch and Speed (2012) and Wang et al. (2017) considered removing unwanted variation or batch effects using negative control genes, which are assumed to be independent of the treatment of interest. In time-series studies of air pollution, Flanders et al. (2011) and Flanders et al. (2017) considered partial correction of residual confounding using a future exposure to air pollution as a negative control exposure. Miao and Tchetgen Tchetgen (2017) extended their method by incorporating both past and future exposures as multiple negative control exposures to further attenuate confounding bias.

The aforementioned methods rely on fairly restrictive assumptions such as rank preservation (Tchetgen Tchetgen, 2013), monotonicity (Sofer et al., 2016), or linear models for the outcome and the unmeasured confounder (Gagnon-Bartsch and Speed, 2012; Wang et al., 2017; Flanders et al., 2011, 2017). In a recent paper, Miao et al. (2018) proposed nonparametric identification of causal effects using a pair of negative control exposure and outcome variables under certain completeness conditions. Their work focused primarily on providing sufficient identification conditions and less so on inference. Ideally, one would in principle aim to obtain inferences in the nonparametric model under which causal effects are identifiable. However, in practice, because one may wish to account for a moderate to large number of observed confounders, nonparametric inference may not be feasible due to the curse of dimensionality.

In this paper, we propose to resolve this difficulty by developing a general semiparametric framework for inferences about the average treatment effect (ATE) in the context of categorical unmeasured confounding adjustment using a pair of negative control exposure and outcome variables while accounting for a possibly large number of observed confounders. In particular, we first extend the identification result of Miao et al. (2018) to a allow for a weaker set of conditions, and provide an alternative representation of the identifying functional for the ATE. The representation is a difference between the standard g-formula of Robins (1986) that fails to account for unmeasured confounding, and an explicit bias correction term that adjusts for unmeasured confounding bias leveraging a pair of negative controls. We then characterize three semiparametric estimators of the ATE that are consistent under three different semiparametric models. Each of the estimators operates on a subset of components of the likelihood for the observed data, and therefore may be severely biased if the corresponding model is misspecified. We carefully combine these strategies into a multiply robust estimator that produces valid inference provided one out of three models is correct, without necessarily knowing which one is indeed correct (Robins et al., 1994; Vansteelandt et al., 2008; Tchetgen Tchetgen and Shpitser, 2012; Rotnitzky et al., 2017). The multiply robust estimator operates on the union of the three semiparametric models and thus offers protection against model misspecification. Furthermore, our proposed multiply robust estimator is locally efficient in the sense that when all working models are correctly specified, our estimator achieves the semiparametric efficiency bound for estimating the ATE under the union model.

The paper is organized as follows. In Section 2 we extend the nonparametric identification results of Miao et al. (2018), and provide an alternative representation of their identifying functional for the ATE, which opens up an opportunity for multiply robust estimation. For ease of exposition, we describe our results in the simple case of binary negative controls and unmeasured confounder in Section 3, where we propose a variety of semiparametric estimators including a multiply robust estimator. We extend our results to the more general setting of polytomous unmeasured confounding and negative controls in Section C of the supplementary material. In Section 4 we assess finite sample performance of our proposed estimators via extensive simulations. We illustrate our methods with an application to the postlicensure surveillance of vaccine safety in Section 5. We close with a brief discussion in Section 6.

2 Identification and reparameterization

We consider estimating the effect of a treatment A on an outcome Y subject to confounding by both observed covariates X and unobserved categorical variables U. Let $Y(a), a = 0, 1$ denote the counterfactual outcome that would be observed if the treatment were a. We are interested in the ATE defined as $E[Y(1) - Y(0)]$. Suppose that we also observe an auxiliary exposure variable Z and an auxiliary outcome variable W, and let $Y(a,z)$ and $W(a,z)$ denote the corresponding counterfactual values that would be observed had the primary treatment and auxiliary exposure taken value (a,z). Then Z and W are negative control exposure and negative control outcome respectively if they satisfy the following assumptions.

Assumption 1. *Negative control exposure:* $Y(a,z) = Y(a)$, *for all* z *almost surely; Negative control outcome:* $W(a,z) = W$ *for all* a, z *almost surely.*

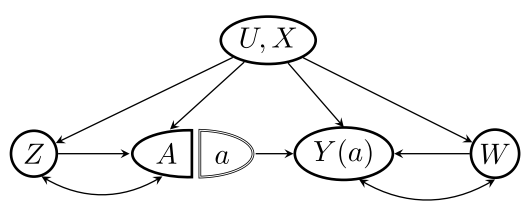
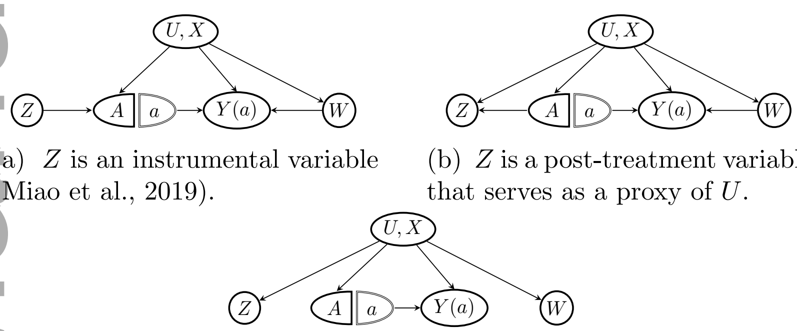


Figure 1: Single world intervention graph with unmeasured confounding U and double negative control Z and W (Richardson and Robins, 2013). The bi-directed arrow between Z and A (Y and W) indicates potential unmeasured common causes of Z and A (Y and W).

Figure 1 presents a single world intervention graph (SWIG, Richardson and Robins (2013)) illustrating an instance of the causal model under consideration. A key assumption satisfied by this graph is the conditional independence assumption stated below, which is required for identification of the causal effect.

Assumption 2. *Latent ignorability:* $(Z, A) \perp\!\!\!\perp (Y(a), W) \mid (U, X)$.

Assumption 2 states that U and X suffice to account for confounding of the relationship between (Z, A) and $(Y(a), W)$, whereas X alone may not. Moreover, U includes all unmeasured common causes of Z, A, Y, and W except for that of the Z-A association and Y-W association. It is important to emphasize that Figure 1 is not the only SWIG that satisfies the negative control assumptions. Figure 2 presents examples of alternative graphs all of which encode Assumption 2. For example, a special case is when Z is an instrumental variable with the additional assumption that $Z \perp\!\!\!\perp U$, as shown in Figure 2a (Miao et al., 2019). Alternatively Z can be a post-treatment variable that serves as a proxy of U, as shown in Figure 2b. Furthermore, Figure 2c presents a scenario where Z and W can be surrogates of U that satisfy the additional assumption that $(Z, W) \perp\!\!\!\perp (A, Y) \mid (U, X)$, which is the nondifferential error assumption (Kuroki and Pearl, 2014). In this scenario, the roles of Z and W can be switched.



(a) Z is an instrumental variable (Miao et al., 2019).

(b) Z is a post-treatment variable that serves as a proxy of U.

(c) Z and W are surrogates of U. Their roles can be switched.

Figure 2: Examples of alternative single world intervention graphs. We suppressed the bi-directed arrow between Z and A (Y and W) because the common causes of Z and A (Y and W) do not confound the Y-A relationship.

Remark 1. *In practice, specification of the unmeasured confounder is helpful for justifying the validity of negative controls. In certain scenarios, however, we do not need to know what U is. For example, an underappreciated causal tenet is that the future does not affect the past. As such, with time series or longitudinal data, future exposure and past outcome may serve as negative control exposure and outcome, respectively, assuming no feedback effect from past outcome to future exposure. In this case, we can control for unmeasured confounders shared over time without singling out a specific U (Miao and Tchetgen Tchetgen, 2017).*

Assumption 3. *Consistency: $Y(a) = Y$ almost surely when $A = a$; Positivity: $0 < P(A = a, Z = z \mid X) < 1$ for all a, z almost surely.*

The consistency assumption ensures that the exposure is defined with enough specificity such that among people with $A = a$, the observed outcome Y is a realization of the potential outcome value $Y(a)$. The positivity assumption states that in all observed covariate strata there are always some individuals with treatment and negative control exposure values $(A = a, Z = z)$, for all a, z.

2.1 Identification with categorical negative control variables

In this paper, we consider the scenario where W, Z, and U are categorical. Suppose W, Z, and U take on $|W|$, $|Z|$, and $|U|$ possible values denoted as w_i, z_j, and u_s, for $i = 0, \dots, |W| - 1$, $j = 0, \dots, |Z| - 1$, and $s = 0, \dots, |U| - 1$ respectively, where $| \cdot |$ denotes the cardinality of a categorical variable. Let $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ denote a $|W| \times |Z|$ matrix with $P(\mathbf{W} \mid \mathbf{Z}, a, x)_{i,j} = P(W = w_{i-1} \mid Z = z_{j-1}, A = a, X = x)$, $P(\mathbf{W} \mid \mathbf{U}, x)$ a $|W| \times |U|$ matrix with $P(\mathbf{W} \mid \mathbf{U}, x)_{i,s} = P(W = w_{i-1} \mid U = u_{s-1}, X = x)$, and $P(\mathbf{U} \mid \mathbf{Z}, a, x)$ a $|U| \times |Z|$ matrix with $P(\mathbf{U} \mid \mathbf{Z}, a, x)_{s,j} = P(U = u_{s-1} \mid Z = z_{j-1}, A = a, X = x)$. Similarly, let $E[Y \mid \mathbf{Z}, a, x]$ denote a $1 \times |Z|$ vector with $E[Y \mid \mathbf{Z}, a, x]_j = E[Y \mid Z = z_{j-1}, A = a, X = x]$, $E[Y \mid \mathbf{U}, a, x]$ a $1 \times |U|$ vector with $E[Y \mid \mathbf{U}, a, x]_s = E[Y \mid U = u_{s-1}, A = a, X = x]$, and $P(\mathbf{W} \mid x)$ a $|W| \times 1$ vector with $P(\mathbf{W} \mid x)_i = P(W = w_{i-1} \mid X = x)$. The following describes a sufficient condition under which the ATE is nonparametrically identified.

Assumption 4. *Both Z and W have at least as many categories as U, i.e., $|Z| \geq |U|$ and $|W| \geq |U|$. Both $P(\mathbf{W} \mid \mathbf{U}, x)$ and $P(\mathbf{U} \mid \mathbf{Z}, a, x)$ are full rank with rank $|U|$ at all values of a and x.*

Remark 2. *Under Assumption 4, $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ has rank $|U|$, which is proved in Section A of the supplementary material. Therefore, one can infer $|U|$ from the rank of $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ (Choi et al., 2017).*

Assumption 4 imposes requirements on candidate negative controls for identification. Intuitively, both Z and W serve as proxies of U. Therefore, they should have at least as many possible values as U. They should also be strongly associated with U such that variation in U can be recovered from variation in Z and W. This is reflected by the requirement that the columns of $P(\mathbf{W} \mid \mathbf{U}, x)$ and the rows of $P(\mathbf{U} \mid \mathbf{Z}, a, x)$ must be linearly independent vectors. In practice, it is recommended to collect a negative control variable with a rich set of possible levels, or multiple negative control variables that can be combined into a composite negative control with as many categories as possible. However, selection of valid negative control variable must be based on reliable subject matter knowledge because Assumptions 1-4 must be met.

The following lemma demonstrates identification of $E[Y(a)]$, which is proved in Section A of the supplementary material.

Lemma 1. *Under Assumptions 1 – 4, there exist a $1 \times |W|$ vector $h(a,x)$ such that*

$$E[Y \mid \mathbf{Z}, a, x] = h(a,x)P(\mathbf{W} \mid \mathbf{Z}, a, x), \tag{1}$$

and $E[Y(a)]$ is nonparametrically identified by $E[Y(a)] = \int_{\mathcal{X}} h(a,x)P(\mathbf{W} \mid x)f(x)dx$, where $f(x)$ denotes the density function of X. Therefore, the ATE, denoted as Δ, is uniquely identified by

$$\Delta = \int_{\mathcal{X}} [h(1,x) - h(0,x)]P(\mathbf{W} \mid x)f(x)dx. \tag{2}$$

As stated in Remark 2, $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ has rank $|U|$ under Assumption 4. When $|Z| = |W| = |U|$, $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ is full rank and the linear system (1) has a unique solution

$$h(a,x) = E[Y \mid \mathbf{Z}, a, x]P(\mathbf{W} \mid \mathbf{Z}, a, x)^{-1}. \tag{3}$$

Therefore, Lemma 1 implies the identification result of Miao et al. (2018) under a stronger assumption that $|Z| = |W| = |U|$, which is stated in the following corollary.

Assumption 4′. *Completeness: $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ is invertible with $|Z| = |W| = |U| = k+1$, $k \geq 0$.*

Corollary 1. *Under Assumptions 1 – 3 and 4′, $E[Y(a)]$ is nonparametrically identified by*

$$E[Y(a)] = \int_{\mathcal{X}} E[Y \mid \mathbf{Z}, a, x]P(\mathbf{W} \mid \mathbf{Z}, a, x)^{-1}P(\mathbf{W} \mid x)f(x)dx.$$

Therefore, the ATE is given by

$$\begin{aligned} \Delta = &\int_{\mathcal{X}} E[Y \mid \mathbf{Z}, A=1, X=x]P(\mathbf{W} \mid \mathbf{Z}, A=1, X=x)^{-1}P(\mathbf{W} \mid X=x)f(x)dx \\ -&\int_{\mathcal{X}} E[Y \mid \mathbf{Z}, A=0, X=x]P(\mathbf{W} \mid \mathbf{Z}, A=0, X=x)^{-1}P(\mathbf{W} \mid X=x)f(x)dx. \end{aligned} \tag{4}$$

When $|Z| > |U|$ or $|W| > |U|$, $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ is rank deficient with linearly dependent rows or columns. In this case, there are infinite solutions to the linear system (1). Nevertheless, $E[Y(a)]$ remains uniquely identified. Note that there always exists an invertible $|U| \times |U|$ submatrix of $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ formed by deleting $|W| - |U|$ rows or $|Z| - |U|$ columns of $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ (Gómez et al., 2008). The $|W| - |U|$ rows or $|Z| - |U|$ columns correspond to free levels in W or Z that are redundant for identification but may improve efficiency.

We propose two strategies for estimation of Δ when $|Z| > |U|$ or $|W| > |U|$. One is to obtain a maximum likelihood estimator of $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ and its Moore-Penrose inverse denoted as $P(\mathbf{W} \mid \mathbf{Z}, a, x)^+$. A particular solution to (1) is given by $h(a, x) = E[Y \mid \mathbf{Z}, a, x]P(\mathbf{W} \mid \mathbf{Z}, a, x)^+$. In fact, by Theorem 2 of James (1978), the complete set of solutions to (1) is given by $h(a, x) = E[Y \mid \mathbf{Z}, a, x]P(\mathbf{W} \mid \mathbf{Z}, a, x)^+ + \tau(a, x)^{\intercal}[\mathbb{I} - P(\mathbf{W} \mid \mathbf{Z}, a, x)P(\mathbf{W} \mid \mathbf{Z}, a, x)^+]$, as $\tau(a, x)$, a vector function, varies over all possible values in $\{f : (a, x) \to R^{|W|}\}$. The second is to coarsen levels in Z and W until the coarsened variables satisfy Assumption 4′ (Kuroki and Pearl, 2014; Miao et al., 2018). Suppose there are m possible sets of coarsened negative control variables, then an estimator can be obtained by the generalized method of moments, i.e., $\hat{\Delta} = \arg\min_{\Delta}[\mathbb{P}_n\hat{g}(\Delta)]^{\intercal}\hat{W}[\mathbb{P}_n\hat{g}(\Delta)]$, where $\hat{g}(\Delta)$ is an m-vector with each entry an estimating equation based on an estimated influence function of Δ under a given parametric, semiparametric, or nonparametric model for a given set of coarsened negative control variables, and $\hat{W} = \mathbb{P}_n[\hat{g}(\Delta)\hat{g}(\Delta)^{\intercal}]^{-1}$. Such influence functions are derived in Section 3.

2.2 Reparameterization of Δ for multiply robust estimation

In this section, we provide an alternative parameterization of Δ which opens up an opportunity for multiply robust estimation in the case where $|Z| = |W| = |U| = k + 1$. When $|Z| > |U|$ or $|W| > |U|$, in order to leverage the reparameterization, we use the second strategy described above in Section 2.1, with $g(\Delta)$ being the efficient influence function detailed in Theorem 1 of Section 3.2.

2.2.1 Motivation for multiply robust estimation

As discussed in Section 1, nonparametric estimation of Δ may not be feasible when X is high dimensional or when Z and W have many levels, in which case one may need to resort to estimation under dimension-reducing working models $E[Y \mid \mathbf{Z}, A, X; \theta_1]$, $P(\mathbf{W} \mid \mathbf{Z}, A, X; \theta_2)$, and $P(\mathbf{W} \mid X; \theta_3)$ where θ_1, θ_2, and θ_3 are finite dimensional, resolving the curse of dimensionality. Under such specification of a model for the conditional distribution $P(Y, W, Z, A \mid X; \theta_1, \theta_2, \theta_3)$, one could in principle estimate Δ using the plug-in estimator, which entails estimating θ_1, θ_2, and θ_3 by standard maximum likelihood estimation (MLE) and substituting estimated parameters in Eq. (2) or (4), with the cumulative distribution function of X estimated by the empirical distribution. This is essentially the approach suggested by Miao et al. (2018). However, these working models are not in themselves of scientific interest and may be prone to model misspecification. The plug-in estimator may be severely biased if any of the three models is incorrect.

To resolve this difficulty, we develop a robust inferential approach grounded in semiparametric theory (Bickel et al., 1993; Newey, 1990; Van der Vaart, 1998), detailed in Section 3. Specifically, we consider the task of estimating the functional Δ without any restriction on the observed data distribution, that is estimation in the nonparametric model denoted as $\mathcal{M}_{\text{nonpar}}$. We characterize the efficient influence function (EIF) for Δ in $\mathcal{M}_{\text{nonpar}}$. We then take the EIF as an estimating equation to obtain an estimator of Δ. Similar to the plug-in estimator, the EIF-based estimation entails estimating the distribution of the observed data under a parametric (or semiparametric) working model and then evaluating the EIF under such working model. However, unlike the plug-in estimator, we establish that our EIF-based estimator of Δ remains consistent and asymptotically normal (CAN) even when the observed data likelihood is partially misspecified. In fact, we establish the multiply robust property of our proposed estimator: it remains CAN under the union of three large semiparametric models, each of which restricts a subset of components of the likelihood, allowing the remaining likelihood components to be unrestricted and hence robust to misspecification.

2.2.2 Reparameterization and intuition for identification

An essential step towards constructing our multiply robust estimator involves a careful reparameterization of the functional Δ in terms of variation independent components of the likelihood, such that (mis)specification of one particular component does not impose any restriction on the other components. To this end, we define the following contrasts measuring the observed effects of Z on Y and W at any value (a, x) as

$$\xi_{z_j}^{w_i}(a,x) = P(W = w_i \mid A = a, Z = z_j, X = x) - P(W = w_i \mid A = a, Z = z_0, X = x), i, j = 1, \dots, k;$$
$$\xi_{z_j}^{Y}(a,x) = E[Y \mid A = a, Z = z_j, X = x] - E[Y \mid A = a, Z = z_0, X = x], j = 1, \dots, k,$$

respectively, where z_0 is a user-specified reference level for Z. Likewise, the observed effects of A on Y and W at any values (z, x) are

$$\delta_A^{w_i}(z,x) = P(W = w_i \mid A = 1, Z = z, X = x) - P(W = w_i \mid A = 0, Z = z, X = x), i = 1, \dots, k;$$
$$\delta_A^{Y}(z,x) = E[Y \mid A = 1, Z = z, X = x] - E[Y \mid A = 0, Z = z, X = x],$$

respectively. In addition, we let

$\boldsymbol{\delta_A^W}(z,x) = \{\delta_A^{w_1}(z,x), \delta_A^{w_2}(z,x), \dots, \delta_A^{w_k}(z,x)\}^\intercal$ denote a $k \times 1$ vector;

$\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{Y}}(a,x) = \{\xi_{z_1}^{Y}(a,x), \xi_{z_2}^{Y}(a,x), \ldots, \xi_{z_k}^{Y}(a,x)\}^{\intercal}$ denote a $k \times 1$ vector;

$\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)$ denote a $k \times k$ matrix with $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)_{i,j} = \xi_{z_j}^{w_i}(a,x)$, $i,j = 1,\ldots,k$.

To avoid over-parameterization, we omitted w_0 and z_0 in the contrasts, which are user-specified reference levels for W and Z, respectively. The following lemma gives our alternative representation, which we prove in Section B of the supplementary material.

Lemma 2. *Under Assumptions 1 – 3 and 4′, $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)$ is invertible and Δ in Eq. (4) admits the alternative representation*

$$\begin{aligned} \Delta =& \Delta_{confounded} - \Delta_{bias}, \\ \Delta_{confounded} =& E[\delta_A^Y(Z,X)],\ \Delta_{bias} = E[\boldsymbol{R}(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)], \end{aligned} \tag{5}$$

where $\boldsymbol{R}(a,x) = \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{Y}}(a,x)^{\intercal}\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)^{-1}$ is a $1 \times k$ vector. In addition, Δ_{bias}=0 if there is no unmeasured confounding.

The alternative representation illustrates the intuition behind identification of Δ. In Eq. (5), $\Delta_{\text{confounded}}$ is the standard g-formula which fails to adjust for unmeasured confounding, and Δ_{bias} is a bias correction term which accounts for unmeasured confounding. We note that Δ_{bias} is a scaled version of the observed association between A and W. In fact, by Assumptions 1 and 2, $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)$ should be zero if there is no unmeasured confounding, and thus a nonzero $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)$ captures confounding bias. The scaling factor $\boldsymbol{R}(1-A,X)$ accounts for the fact that the effect of U on Y may not be on the same scale as the effect of U on W, and therefore the bias captured by $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)$ needs to be carefully rescaled. To identify the ratio of the effects of U on Y and U on W, we note that conditional on A and X, any association between Z and Y (Z and W) is governed by the effect of U on Y (U on W). Therefore the ratio of the observed Z effects, i.e., $\boldsymbol{R}(1-A,X)$, recovers the ratio of the unobserved U effects. We further illustrate the intuition behind identification and reparameterization with an example in Section B.1 of the supplementary material.

Decomposition of the causal effect estimand into the standard g-formula and an explicit bias correction term simplifies our inferential task, because semiparametric estimation of $\Delta_{\text{confounded}}$ has been extensively studied (Robins et al., 1994; Robins, 2000; Scharfstein et al., 1999; Van der Laan and Robins, 2003; Bang and Robins, 2005; Tan, 2006; Tsiatis, 2007). Therefore we mainly study robust estimation of Δ_{bias}, which together with $\Delta_{\text{confounded}}$ provides robust estimation of the ATE.

For ease of exposition, in the following sections we develop our semiparametric approach in the setting where W, Z, and U are binary variables. We extend our results to general settings with polytomous W, Z, and U in Section C of the supplementary material.

3 Semiparametric estimation in the binary case

When Z, W, U are binary, i.e., $k = 1$, $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(z,x)$, $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{Y}}(a,x)$, $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)$, and $\boldsymbol{R}(a,x)$ simplify to the following scalar functions

$$\begin{aligned}
\delta_A^W(z,x) =& E[W \mid A=1, Z=z, X=x] - E[W \mid A=0, Z=z, X=x],\\
\xi_Z^Y(a,x) =& E[Y \mid A=a, Z=1, X=x] - E[Y \mid A=a, Z=0, X=x],\\
\xi_Z^W(a,x) =& E[W \mid A=a, Z=1, X=x] - E[W \mid A=a, Z=0, X=x],\\
R(a,x) =& \frac{\xi_Z^Y(a,x)}{\xi_Z^W(a,x)},
\end{aligned} \tag{6}$$

and representation of Δ in Eq. (5) is accordingly simplified. Note that careful specification of $R(A,X)$, $\xi_Z^W(A,X)$, and $\xi_Z^Y(A,X)$ is critical as they are in general not variation independent; that is, model specification for $R(A,X)$ and $\xi_Z^W(A,X)$ would imply a model for $\xi_Z^Y(A,X)$.

3.1 Working models and three classes of semiparametric estimators

We now formally introduce variation independent components of the observed data likelihood for estimation of Δ to facilitate robust estimation. First, we note that the mean of W given A, Z, and X can be written as

$$E[W \mid A, Z, X] = E[W \mid A=0, Z=0, X] + \xi_Z^W(A=0,X)Z + \delta_A^W(Z=0,X)A + \eta_{AZ}^W(X)AZ, \tag{7}$$

where $\eta_{AZ}^W(\cdot)$ is the additive interaction of A and Z given X with

$$\eta_{AZ}^W(X)AZ = [\xi_Z^W(A,X) - \xi_Z^W(A=0,X)]Z = [\delta_A^W(Z,X) - \delta_A^W(Z=0,X)]A. \tag{8}$$

Furthermore, it is straightforward to verify that

$$E[Y \mid Z, A, X] = E[Y \mid Z=0, A, X] + R(A,X)\xi_Z^W(A,X)Z, \tag{9}$$

which implies that

$$\delta_A^Y(Z,X) = \left[E[Y \mid Z=0, A=1, X] + R(A=1,X)\xi_Z^W(A=1,X)Z\right] - \left[E[Y \mid Z=0, A=0, X] + R(A=0,X)\xi_Z^W(A=0,X)Z\right]. \tag{10}$$

Multiply robust estimation requires positing working models for the following quantities: $E[Y \mid Z=0, A, X]$, $E[W \mid A=0, Z=0, X]$, $\xi_Z^W(A=0,X)$, $\delta_A^W(Z=0,X)$, $\eta_{AZ}^W(X)$, $R(A,X)$, and $f(A,Z \mid X)$, where $f(A,Z \mid X)$ is the joint density of A and Z conditional on X. As X may be high-dimensional and Z and W may have many levels, dimension-reducing parametric (or semi-parametric) working models are used to avoid the curse of dimensionality in practice. Clearly, these working models are not in themselves of scientific interest. Estimators relying on a subset of these models may be biased when the corresponding models are misspecified.

In order to motivate and clarify our doubly robust estimator, we introduce three classes of semiparametric estimators of Δ, which are CAN under the following working models respectively with finite-dimensional indexing parameters:

$\mathcal{M}_1$: Working models $f(A,Z \mid X;\alpha^{A,Z})$ and $R(A,X;\beta^R)$ are correctly specified.

$\mathcal{M}_2$ Working models $f(A,Z \mid X;\alpha^{A,Z})$, and $\xi_Z^W(A,X;\beta^{WZ})$ and $\delta_A^W(Z,X;\beta^{WA})$ satisfying restriction (8) are correctly specified. The interaction model $\eta_{AZ}^W(X;\beta^{WAZ})$ is indexed by β^{WAZ}, which is a sub-vector shared by β^{WZ} and β^{WA}.

$\mathcal{M}_3$: Working models $R(A,X;\beta^R)$, and $E[Y \mid Z=0, A, X;\beta^Y]$ and $E[W \mid A, Z, X;\beta^W]$ with $\beta^W = (\beta^{W0}, \beta^{WZ}, \beta^{WA})$ are correctly specified, where $E[W \mid A, Z, X;\beta^W]$ is parameterized by Eq. (7) and β^{W0} denotes the sub-vector of β^W that indexes the baseline $E[W \mid A=0, Z=0, X]$.

Compared to the full list of variation independent components, we can see that in $\mathcal{M}_1$, $E[Y \mid Z=0, A, X]$, $E[W \mid A=0, Z=0, X]$, $\xi_Z^W(A=0,X)$, $\delta_A^W(Z=0,X)$, and $\eta_{AZ}^W(X)$ are unrestricted; in $\mathcal{M}_2$, $R(A,X)$, $E[Y \mid Z=0, A, X]$, and $E[W \mid A=0, Z=0, X]$ are unrestricted; while in model $\mathcal{M}_3$, $f(A,Z \mid X)$ is unrestricted.

We now describe three semiparametric estimators which are CAN under $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, respectively. Let $\gamma_i, i=1,\ldots,3$ denote the collection of indexing parameters in the corresponding semiparametric working model $\mathcal{M}_i$, which can be estimated under $\mathcal{M}_i$ as detailed in Appendix A.1

of the main manuscript. Let $\hat{\gamma}_i$ denote the estimated parameters, we have

$$\hat{\Delta}_1 = \mathbb{P}_n\left\{\frac{(2A-1)Y}{f(A \mid Z, X; \hat{\gamma}_1)}\right\} - \mathbb{P}_n\left\{E[R(1-A, X) \mid Z, X; \hat{\gamma}_1]\frac{(2A-1)W}{f(A \mid Z, X; \hat{\gamma}_1)}\right\}$$
$$\hat{\Delta}_2 = \mathbb{P}_n\left\{\frac{(2A-1)Y}{f(A \mid Z, X; \hat{\gamma}_2)}\right\} - \mathbb{P}_n\left\{\frac{(2Z-1)Y}{f(Z \mid A, X; \hat{\gamma}_2)}\frac{E[\delta_A^W(Z, X) \mid 1-A, X; \hat{\gamma}_2]}{\xi_Z^W(A, X; \hat{\gamma}_2)}\frac{f(1-A \mid X; \hat{\gamma}_2)}{f(A \mid X; \hat{\gamma}_2)}\right\}$$
$$\hat{\Delta}_3 = \mathbb{P}_n\left\{E[Y \mid A=1, Z, X; \hat{\gamma}_3] - E[Y \mid A=0, Z, X; \hat{\gamma}_3]\right\} - \mathbb{P}_n\left\{R(1-A, X; \hat{\gamma}_3)\delta_A^W(Z, X; \hat{\gamma}_3)\right\},$$

where $\mathbb{P}_n$ is the empirical average operator, i.e., $\mathbb{P}_n(V) = \frac{1}{n}\sum_{i=1}^n V_i$.

Each of the three estimators above may be severely biased if their corresponding model $\mathcal{M}_1$, $\mathcal{M}_2$, or $\mathcal{M}_3$ is misspecified. For example, $\hat{\Delta}_1$ and $\hat{\Delta}_2$ will generally fail to be consistent if $f(A \mid Z, X)$ is misspecified, even if the rest of the components of the likelihood is correctly specified. Therefore, it is critical to develop a multiply robust estimator that remains CAN provided that one, but not necessarily more than one of models $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ is correctly specified, without necessarily knowing which one is indeed correct.

3.2 Efficient influence function in the nonparametric model

We aim to construct an estimator that is CAN under the union model $\mathcal{M}_{\text{union}} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$. To this end, we first characterize the EIF for Δ in the nonparametric model $\mathcal{M}_{\text{nonpar}}$ which does not impose any restriction on the observed data distribution. We then use the EIF as an estimating equation and evaluate it under a working model to obtain an estimator of Δ. We establish multiple robustness and asymptotic normality of this estimator. We also provide a consistent estimator of the asymptotic variance for the proposed estimators.

It is well know that the efficient influence function of $\Delta_{\text{confounded}}$ in $\mathcal{M}_{\text{nonpar}}$ (Robins et al., 1994) is

$$EIF_{\Delta_{\text{confounded}}} = \frac{2A-1}{f(A \mid Z, X)}\big(Y - E[Y \mid A, Z, X]\big) + \big(E[Y \mid A=1, Z, X] - E[Y \mid A=0, Z, X]\big) - \Delta_{\text{confounded}}. \tag{11}$$

In the theorem below, we derive the efficient influence function of Δ_{bias} in $\mathcal{M}_{\text{nonpar}}$, which is combined with $EIF_{\Delta_{\text{confounded}}}$ to obtain the efficient influence function of Δ. Theorem 1 is proved in Section D of the supplementary material.

Theorem 1. *Under Assumptions 1 – 3 and 4′, the efficient influence function of the bias correction* *term Δ_{bias} in the nonparametric model $\mathcal{M}_{nonpar}$ is*

$$\begin{aligned} EIF_{\Delta_{bias}} =& E[R(1-A,X) \mid Z,X] \frac{2A-1}{f(A \mid Z,X)} \Big(W - E[W \mid A,Z,X]\Big) \\ &+ \frac{2Z-1}{f(Z \mid A,X)} \Big(Y - E[Y \mid Z,A,X]\Big) \frac{E[\delta_A^W(Z,X) \mid 1-A,X]}{\xi_Z^W(A,X)} \frac{f(1-A \mid X)}{f(A \mid X)} \\ &+ R(1-A,X)\delta_A^W(Z,X) - \Delta_{bias}. \end{aligned}$$

The efficient influence function of Δ is given by

$$EIF_\Delta(O) = EIF_{\Delta_{confounded}} - EIF_{\Delta_{bias}},$$

where $O = (Y,A,Z,W,Z)$ denotes the observed data. The semiparametric efficiency bound for estimating the ATE in $\mathcal{M}_{nonpar}$ is $E[EIF_\Delta(O)^2]^{-1}$.

Remark 3. *Theorem 1 implies that in $\mathcal{M}_{nonpar}$, all regular and asymptotically linear estimators $\hat{\Delta}$ are asymptotically equivalent and efficient with $\sqrt{n}(\hat{\Delta} - \Delta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n EIF_\Delta(O_i) + o_p(1)$ (Bickel et al., 1993).*

3.3 Multiply robust estimation of Δ

In this section, we consider the scenario where estimation under $\mathcal{M}_{\text{nonpar}}$ is not feasible due to potentially large number of measured covariates, and proceed to estimation under $\mathcal{M}_{\text{union}}$. Specifically, we construct a multiply robust and locally efficient estimator of Δ by taking $EIF_\Delta(O)$ as an estimating equation and evaluating it under a working model for the observed data distribution to solve for Δ. Let

$$\theta = \{(\alpha^{A,Z})^\intercal, (\beta^Y)^\intercal, (\beta^{W0})^\intercal, (\hat{\beta}^{WA})^\intercal, (\hat{\beta}^{WZ})^\intercal, (\hat{\beta}^R)^\intercal\}^\intercal$$

denote the nuisance parameters of the working models in $\mathcal{M}_{\text{union}}$. We estimate θ as the solution of the following collection of estimating equations.

First, we define the following score functions for maximum likelihood estimation of $f(A,Z \mid X;\alpha^{A,Z})$, $E[Y \mid A, Z=0, X; \beta^Y]$, and $E[W \mid A=0, Z=0, X; \beta^{W0}]$

$$\begin{aligned} U_{\alpha^{A,Z}} =& \frac{\partial}{\partial \alpha^{A,Z}} \log f(A,Z \mid X; \alpha^{A,Z}); \\ U_{\beta^Y} =& \frac{\partial}{\partial \beta^Y} \mathbb{1}(Z=0) \log f(Y \mid A, Z=0, X; \beta^Y); \\ U_{\beta^{W0}} =& \frac{\partial}{\partial \beta^{W0}} \mathbb{1}(A=0,Z=0) \log f(W \mid A=0, Z=0, X; \beta^{W0}), \end{aligned}$$

where $f(A,Z \mid X;\alpha^{A,Z})$ is the conditional likelihood of (A,Z), $f(Y \mid A,Z=0,X;\beta^{Y})$ is the conditional likelihood of Y restricted to the subsample with $Z=0$, and $f(W \mid A=0,Z=0,X;\beta^{W0})$ is the conditional likelihood of W restricted to the subsample with $A=0,Z=0$.

Second, because $\delta_A^W(Z,X;\beta^{WA})$, $\xi_Z^W(A,X;\beta^{WZ})$, and $R(A,X;\beta^R)$ do not by themselves give rise to a likelihood function, we estimate them by constructing the following doubly robust g-estimation equations constructed under the union model $\mathcal{M}_{\text{union}}$

$$U_{\beta^{WA},\beta^{WZ}} = \Big(g_0(A,Z,X) - E[g_0(A,Z,X) \mid X;\alpha^{A,Z}]\Big)\Big(W - E[W \mid A,Z,X;\beta^{W0},\beta^{WZ},\beta^{WA}]\Big)$$
$$U_{\beta^R;\beta^Y,\beta^{W0},\beta^{WA}} = \Big(g_1(A,Z,X) - E[g_1(A,Z,X) \mid A,X;\alpha^{A,Z}]\Big)\Big(Y - E[Y \mid Z,A,X;\beta^R,\beta^Y,\beta^{W0},\beta^{WA}]\Big),$$

where $g_0(A,Z,X)$ and $g_1(A,Z,X)$ are user-specified vector functions; $E[g_0(A,Z,X) \mid X;\alpha^{A,Z}]$ and $E[g_1(A,Z,X) \mid X;\alpha^{A,Z}]$ are evaluated under $f(A,Z \mid X;\alpha^{A,Z})$; $E[W \mid A,Z,X;\beta^{W0},\beta^{WZ},\beta^{WA}]$ and $E[Y \mid Z,A,X;\beta^R,\beta^Y,\beta^{W0},\beta^{WA}]$ are parameterized as in (7)-(10). Let $\dim(v)$ denote the length of a vector v. We require that $g_0(A,Z,X)$ is of dimension $\dim(\beta^{WA})+\dim(\beta^{WZ})-\dim(\beta^{WAZ})$, and $g_1(A,Z,X)$ is of dimension $\dim(\beta^R)$ to generate adequate number of estimating equations.

In summary, let

$$U_\theta(O;\theta) = (U^\intercal_{\alpha^{A,Z}}, U^\intercal_{\beta^Y}, U^\intercal_{\beta^{W0}}, U^\intercal_{\beta^{WA},\beta^{WZ}}, U^\intercal_{\beta^R})^\intercal$$

denote the collection of the above defined estimating equations. We estimate θ by solving $\mathbb{P}_n\Big\{U_\theta(\theta)\Big\}=0$, and we denoted the estimator as

$$\hat{\theta} = \{(\hat{\alpha}^{A,Z}_{\text{mle}})^\intercal, (\hat{\beta}^{Y}_{\text{mle}})^\intercal, (\hat{\beta}^{W0}_{\text{mle}})^\intercal, (\hat{\beta}^{WA}_{\text{dr}})^\intercal, (\hat{\beta}^{WZ}_{\text{dr}})^\intercal, (\hat{\beta}^{R}_{\text{dr}})^\intercal\}^\intercal.$$

In particular, $\hat{\beta}^{WA}_{\text{dr}}$ and $\hat{\beta}^{WZ}_{\text{dr}}$ are CAN under the union model $\mathcal{M}_2 \cup \mathcal{M}_3$, and $\hat{\beta}^{R}_{\text{dr}}$ is CAN under the union model $\mathcal{M}_1 \cup \mathcal{M}_3$ (Robins and Rotnitzky, 2001; Wang and Tchetgen Tchetgen, 2018), which is proved in Section E of the supplementary material. We obtain the estimated working models by plugging in $\hat{\theta}$ to equations (7)-(10), which is detailed in Appendix A.2.

The proposed multiply robust estimator solves $\mathbb{P}_n\Big\{EIF_\Delta(O;\Delta,\hat{\theta})\Big\}=0$, where $EIF_\Delta(O;\Delta,\hat{\theta})$ is equal to $EIF_\Delta(O)$ evaluated at $(\Delta,\hat{\theta})$. That is, the multiply robust estimator is

$$\hat{\Delta}_{\text{mr}} = \hat{\Delta}_{\text{confounded,mr}} - \hat{\Delta}_{\text{bias,mr}}$$

where

$$\hat{\Delta}_{\text{confounded,mr}} = \mathbb{P}_n\Big\{\frac{2A-1}{f(A \mid Z, X; \hat{\theta})}\big(Y - E[Y \mid A, Z, X; \hat{\theta}]\big) + \big(E[Y \mid A=1, Z, X; \hat{\theta}] - E[Y \mid A=0, Z, X; \hat{\theta}]\big)\Big\}$$

$$\begin{aligned}\hat{\Delta}_{\text{bias,mr}} = \mathbb{P}_n\Big\{ & E[R(1-A, X) \mid Z, X; \hat{\theta}]\frac{2A-1}{f(A \mid Z, X; \hat{\theta})}\Big(W - E[W \mid A, Z, X; \hat{\theta}]\Big) \\ & + \frac{2Z-1}{f(Z \mid A, X; \hat{\theta})}\Big(Y - E[Y \mid A, Z, X; \hat{\theta}]\Big)\frac{E[\delta_A^W(Z, X) \mid 1-A, X; \hat{\theta}]}{\xi_Z^W(A, X; \hat{\theta})}\frac{f(1-A \mid X; \hat{\theta})}{f(A \mid X; \hat{\theta})} \\ & + R(1-A, X; \hat{\theta})\delta_A^W(Z, X; \hat{\theta})\Big\}.\end{aligned}$$

The multiply robust estimator combines three semiparametric estimation strategies to produce robust inference provided one out of three working models is correct, without necessarily knowing which one is indeed correct. This can be seen by the fact that each of the three semiparametric estimators $\hat{\Delta}_i$ can be obtained by setting the components unrestricted in $\mathcal{M}_i$ to zero in the above multiply robust estimator. Specifically, $\hat{\Delta}_1$ can be obtained by setting $E[Y \mid Z=0, A, X]$, $E[W \mid A=0, Z=0, X]$, $\xi_Z^W(A=0, X)$, $\delta_A^W(Z=0, X)$, and $\eta_{AZ}^W(X)$ to zero, $\hat{\Delta}_2$ can be obtained by setting $E[Y \mid Z=0, A, X]$, $E[W \mid A=0, Z=0, X]$, and $R(A, X)$ to zero, and $\hat{\Delta}_3$ can be obtained by setting $1/f(A \mid Z, X)$ and $1/f(Z \mid A, X)$ to zero. In particular, the multiply robust estimator of $\Delta_{\text{bias}} = E[R(1-A, X)\delta_A^W(Z, X)]$ does not require correct specification of both $R(1-A, X)$ and $\delta_A^W(Z, X)$. In fact, we improve robustness by incorporating the propensity of both exposures such that when $f(A, Z \mid X)$ is correctly specified, $\hat{\Delta}_{\text{bias,mr}}$ is consistent if either $R(1-A, X)$ or $\delta_A^W(Z, X)$ is correctly specified. Our proposed estimator is also locally efficient in the sense that when all working models are correctly specified, $\hat{\Delta}_{\text{mr}}$ achieves the semiparametric efficiency bound for estimating Δ in $\mathcal{M}_{\text{union}}$. Theorem 2 below summarizes the multiply robust and locally efficient property of $\hat{\Delta}_{\text{mr}}$.

Theorem 2. *Under Assumptions 1 – 3 and 4′ and standard regularity conditions stated in Section E of the supplementary material, $\sqrt{n}(\hat{\Delta}_{mr} - \Delta)$ is regular and asymptotic linear under $\mathcal{M}_{union}$ with influence function*

$$IF_{union}(O; \Delta, \theta^*) = EIF_\Delta(O; \Delta, \theta^*) - \frac{\partial EIF_\Delta(O; \Delta, \theta)}{\partial \theta^\intercal}\Big|_{\theta^*} E\Big\{\frac{\partial U_\theta(O; \theta)}{\partial \theta^\intercal}\Big|_{\theta^*}\Big\}^{-1} U_\theta(O; \theta^*),$$

and thus $\sqrt{n}(\hat{\Delta}_{mr} - \Delta) \to_d N(0, \sigma_\Delta^2)$, where $\sigma_\Delta^2(\Delta, \theta^*) = E[IF_{union}(O; \Delta, \theta^*)^2]$ and θ^* denotes the probability limit of $\hat{\theta}$. Furthermore, $\hat{\Delta}_{mr}$ is locally semiparametric efficient in the sense that it achieves the semiparametric efficiency bound for Δ in $\mathcal{M}_{union}$ at the intersection submodel

$\mathcal{M}_{intersect} = \mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3$ *where* $\mathcal{M}_1$, $\mathcal{M}_2$, *and* $\mathcal{M}_3$ *are all correctly specified.*

We prove Theorem 2 in Section E of the supplementary material. The rationale behind multiple robustness is based on the following key observation. A multiply robust estimator is bound to exist if one can describe an unbiased estimating equation in each of the submodels that form the union model. It then suffices to show that the multiply robust estimating equation (i.e. the efficient influence function) reduces to each estimating equation under the corresponding submodel of the union model, by setting components which are left unrestricted in the submodel to a singleton value. For inference on Δ, a consistent standard error estimator follows from standard M-estimation theory, which is detailed in Section E.3 of the supplementary material. We implemented the standard error estimator in both simulation and application studies. Alternatively, nonparametric bootstrap may be used in practice, which is justified by the asymptotic linearity of the estimator (Cheng et al., 2010).

4 Simulation study

We investigate the finite sample performance of the various estimators of ATE described in Section 3. We simulate 4000 samples of size $n = 2000$ under the following data generating mechanism

- $X = (X_1, \ldots, X_8, X_7X_8)$ where $X_j \overset{\text{iid}}{\sim} N(0,1), j = 1, \ldots, 8$;
- A is Bernoulli with $P(A = 1 \mid X) = \text{expit}(-0.01 + \alpha^\intercal X)$;
- Z is Bernoulli with $P(Z = 1 \mid A, X) = \text{expit}(-0.01 - 0.2A + \alpha^\intercal X)$;
- U is Bernoulli with $E[U \mid Z, A, X] = 0.4Z + 0.4AZ$;
- W is Bernoulli with $E[W \mid U = 0, X] = \text{expit}(-1 + \beta^\intercal X)$, $E[W \mid U = 1, X] - E[W \mid U = 0, X] = 0.5$;
- Y is Bernoulli with $E[Y \mid A = 0, U = 0, X] = \text{expit}(-1 + \beta^\intercal X)$, $E[Y \mid A, U = 1, X] - E[Y \mid A, U = 0, X] = 0.25A$, and $E[Y \mid A = 1, U, X] - E[Y \mid A = 0, U, X] = 0.25U$,

where $\alpha = -10^{-2} \times (1,1,1,1,1,1,1,1,-20)$ and $\beta = -10^{-1} \times (1,1,1,1,1,1,1,1,1)$. These parameters are chosen to ensure that $\Pr(U = 1 \mid Z, A, X)$, $\Pr(W = 1 \mid U, X)$, and $\Pr(Y = 1 \mid U, X)$ are between 0 and 1. The above models imply

- $\xi_Z^W(A, X) = 0.2 + 0.2A$, $\delta_A^W(Z, X) = 0.2Z$, $E[W \mid Z = 0, A = 0, X] = \text{expit}(-1 + \beta^\intercal X)$;
- $\xi_Z^Y(A, X) = 0.2A$, $\delta_A^Y(Z, X) = 0.2Z$, $E[Y \mid Z = 0, A = 0, X] = \text{expit}(-1 + \beta^\intercal X)$;

- $R(A, X) = 0.5A$.

We evaluate the performance of the following five estimators of the ATE: three semiparametric estimators $\hat{\Delta}_1$, $\hat{\Delta}_2$, and $\hat{\Delta}_3$, the plug-in estimator discussed in Section 2.2.1 which we refer to as the MLE estimator hereafter, and the multiply robust (MR) estimator $\hat{\Delta}_{\mathrm{mr}}$. The true ATE is 0.07 on the risk difference scale. We consider the following scenarios to investigate the impact of modeling error:

- All models are correctly specified;
- $\mathcal{M}_2$ and $\mathcal{M}_3$ are wrong: $E[W \mid A, Z, X]$ is misspecified by assuming that both $\xi_Z^W(A, X)$ and $\delta_A^W(Z, X)$ are constant;
- $\mathcal{M}_1$ and $\mathcal{M}_3$ are wrong: $R(A, X)$ is misspecified by assuming that $R(A, X)$ is a constant;
- $\mathcal{M}_1$ and $\mathcal{M}_2$ are wrong: $f(Z \mid A, X)$ is misspecified by omitting the interaction term X_7X_8;
- All models are wrong: $f(Z \mid A, X)$ and $E[Y \mid A, Z, X]$ are misspecified by omitting the interaction term X_7X_8.

Table 1: Operating characteristics of estimators under different model misspecification scenarios.

Scenario	**Method**	**Bias** ($\times 10^3$)	**Var** ($\times 10^3$)	**Proportion Bias (% ATE)**	**MSE** ($\times 10^3$)	**95% CI Coverage**
All models are correct	Δ_1	-0.46	0.45	-0.65	0.45	0.95
	Δ_2	-0.37	0.62	-0.53	0.62	0.95
	Δ_3	-0.06	0.14	-0.08	0.14	0.95
	MLE	-0.49	0.10	-0.70	0.10	0.95
	MR	-0.39	0.73	-0.55	0.73	0.95
$\mathcal{M}_1$ **correct** $\mathcal{M}_2, \mathcal{M}_3$ **misspecified**	Δ_2	-7.10	0.48	-10.08	0.53	0.94
	Δ_3	-7.10	0.14	-10.08	0.19	0.91
	MLE	-24.05	6.47	-34.15	7.04	0.91
	MR	2.54	0.49	3.61	0.49	0.95
$\mathcal{M}_2$ **correct** $\mathcal{M}_1, \mathcal{M}_3$ **misspecified**	Δ_1	-0.51	0.50	-0.73	0.50	0.94
	Δ_3	-5.04	0.60	-7.22	0.63	0.95
	MR	0.27	0.56	0.39	0.56	0.95
$\mathcal{M}_3$ **correct** $\mathcal{M}_1, \mathcal{M}_2$ **misspecified**	Δ_1	-0.25	0.45	-0.36	0.45	0.95
	Δ_2	-1.22	0.61	-1.75	0.61	0.95
	MR	-0.05	1.14	-0.08	1.14	0.95
All models are misspecified	Δ_1	-0.25	0.45	-0.36	0.45	0.95
	Δ_2	-1.22	0.61	-1.75	0.61	0.95
	Δ_3	-2.80	0.14	-4.01	0.15	0.94
	MLE	-2.15	0.10	-3.08	0.10	0.94
	MR	0.60	1.10	0.86	1.10	0.95

Note: we trimmed 5% tail of the second scenario due to extreme value of the MLE estimates.

Table 1 summarizes the operating characteristics of $\hat{\Delta}_1$, $\hat{\Delta}_2$, $\hat{\Delta}_3$, the MLE estimator, and the MR estimator $\hat{\Delta}_{\mathrm{mr}}$ under the above model misspecification scenarios. We evaluated these estimators in terms of mean bias (scaled by 10^3), variance (scaled by 10^3), bias calculated as the proportion of the true ATE, mean squared error (MSE, scaled by 10^3), and coverage of 95% confidence intervals based on direct standard error estimates. The performance of the MLE estimator is not shown when $R(A, X)$ or $f(Z \mid A, X)$ are misspecified because it does not require specification of $R(A, X)$ or $f(Z \mid A, X)$ and thus remains unchanged under such misspecifications. Our proposed multiply robust estimator remained stable with relatively small bias across all scenarios, although as expected it had slightly larger variability. The multiply robust estimator performs better when all models are misspecified than if $\mathcal{M}_2$ and $\mathcal{M}_3$ are misspecified, which may not be the general case in practice as the theory does not necessarily justify it. In contrast, the MLE estimator and the other three semiparametric estimators that rely on $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ can be substantially biased when their corresponding model was misspecified. The 95% CI coverages were close to the nominal level with correctly specified model which indicated that our proposed standard error estimation provided valid inference. These results confirmed our theoretical results in finite sample and demonstrated the advantages of the proposed multiply robust estimator.

5 Observational postlicensure vaccine safety surveillance

We apply our method to an observational vaccine safety study comparing risk of medically attended fever, a common adverse event following vaccination, among children who received a combination DTaP-IPV-Hib (diphtheria and tetanus toxoids and acellular pertussis adsorbed, inactivated poliovirus, and Haemophilus influenzae type b) vaccine with children who received other DTaP-containing comparator vaccines (Nelson et al., 2013). The study population consisted of children aged 6 weeks to 2 years enrolled at Kaiser Permanente Washington from September 2008 to January 2011. Healthcare databases routinely captured information on demographics, immunizations, and diagnosis of fever within a 5-day post-vaccination risk window based on the International Classification of Diseases, Ninth Revision (ICD-9-CM) code.

In the absence of randomization, causal inference methods can be applied to evaluate the adverse effect of DTaP-IPV-Hib vaccine. However, because such administrative data are not collected for research purposes, potential bias due to unmeasured confounding can undermine the validity of causal conclusion. In particular, parents of infants may request separate injections or the combination vaccine due to unmeasured health-seeking preference, and such health-seeking behavior may be associated with fever diagnosis. To explore the possibility of confounding due to health-seeking behavior, the study monitored presence of injury/trauma (ICD-9 code 800-904, 910-959) and ringworm (ICD-9 code 110) within 30 days post vaccination, which are not expected to be related to the vaccine-outcome pair of interest. In particular, injury/trauma is unlikely to be causally affected by DTaP-IPV-Hib vaccination but may be associated with parents' health-seeking behavior on behalf of their children. Similarly, ringworm is unlikely to be a cause of fever that occurs during the 5-day risk window but may also be associated with health-seeking behavior. Therefore, we take injury/trauma as a negative control outcome and ringworm as a negative control exposure to detect and account for potential unmeasured confounding. During the study, 27,064 DTaP-IPV-Hib vaccinations were administered, among which 60 fevers (0.22%) were observed within the risk window. In contrast, 19,677 comparator vaccines were administered with 46 fevers (0.23%) observed. There were 45 ringworm cases and 46 injury/trauma cases. Sex and age group at vaccination (<5 months or 5 months$-$2 years) were also recorded.

Because A, Z, and X are all binary, nonparametric (NP) estimation based on empirical frequencies is in fact feasible. This is done by fitting a saturated model for each component of the likelihood by including main effects and all possible interactions. For example, the negative control outcome model was specified as $E[W \mid A, Z, X_1, X_2] = \alpha_0 + \alpha_A A + \alpha_Z Z + \alpha_{X_1} X_1 + \alpha_{X_2} X_2 + \alpha_{A:Z} AZ + \alpha_{A:X_1} AX_1 + \alpha_{Z:X_1} ZX_1 + \alpha_{A:X_2} AX_2 + \alpha_{Z:X_2} ZX_2 + \alpha_{X_1:X_2} X_1 X_2 + \alpha_{A:Z:X_1} AZX_1 + \alpha_{A:Z:X_2} AZX_2 + \alpha_{A:X_1:X_2} AX_1X_2 + \alpha_{Z:X_1:X_2} ZX_1X_2 + \alpha_{A:Z:X_1:X_2} AZX_1X_2$, where X_1 denotes age group and X_2 denotes sex. As stated in Remark 3, under $\mathcal{M}_{\text{nonpar}}$ when all nuisance parameters were nonparametrically estimated, all methods should produce exactly the same point estimate and confidence interval. We thus took the NP model as the true model to illustrate robustness to departure from the NP model via model restrictions in the following scenarios

- $\mathcal{M}_2$ and $\mathcal{M}_3$ are restricted: $E[W \mid A, Z, X]$ is fitted without age-sex interaction;
- $\mathcal{M}_1$ and $\mathcal{M}_3$ are restricted: $R(A, X)$ is fitted without age-sex interaction;
- $\mathcal{M}_1$ and $\mathcal{M}_2$ are restricted: $f(Z \mid A, X)$ is fitted without age-sex interaction;
- All are restricted: $E[W \mid A, Z, X]$ and $R(A, X)$ are fitted without age-sex interaction.

Table 2 lists for each method the point estimates (scaled by 10^3) of Δ, $\Delta_{\text{confounded}}$, Δ_{bias} and their 95% confidence intervals (scaled by 10^3), the bias evaluated as the proportion of the ATE under the NP model which is taken as the true value, and the p-value from a Wald-test of

Table 2: Adverse effect of DTaP-IPV-Hib vaccine on fever among children.

Scenario	Method	$\hat{\Delta}$ (95% CI)	Prop Bias	p-val	$\hat{\Delta}_{\text{confounded}}$ (95% CI)	$\hat{\Delta}_{\text{bias}}$ (95% CI)
All models are NP	Δ_1	1.7 (-1.1, 4.5)	0.0	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.3)
	Δ_2	1.7 (-1.1, 4.5)	0.0	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.3)
	Δ_3	1.7 (-1.1, 4.5)	0.0	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.3)
	MLE	1.7 (-1.1, 4.5)	0.0	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.3)
	MR	1.7 (-1.1, 4.5)	0.0	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.3)
$\mathcal{M}_1$ **is NP** $\mathcal{M}_2, \mathcal{M}_3$ **are restricted**	Δ_2	1.7 (-1.4, 4.8)	-0.3	0.3	0.5 (-0.5, 1.4)	-1.2 (-4.1, 1.6)
	Δ_3	0.5 (-0.5, 1.4)	-72.3	0.3	0.5 (-0.5, 1.4)	-0.0 (-0.3, 0.3)
	MLE	1.6 (-1.5, 4.7)	-5.6	0.3	0.5 (-0.5, 1.4)	-1.1 (-4.0, 1.7)
	MR	1.6 (-1.3, 4.5)	-5.2	0.3	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.5)
$\mathcal{M}_2$ **is NP** $\mathcal{M}_1, \mathcal{M}_3$ **are restricted**	Δ_1	1.7 (-1.1, 4.4)	-3.1	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.7, 1.3)
	Δ_3	1.7 (-1.1, 4.4)	-3.1	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.7, 1.3)
	MR	1.7 (-1.1, 4.6)	1.6	0.2	0.5 (-0.5, 1.4)	-1.3 (-3.9, 1.3)
$\mathcal{M}_3$ **is NP** $\mathcal{M}_1, \mathcal{M}_2$ **are restricted**	Δ_1	1.7 (-1.1, 4.5)	0.0	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.3)
	Δ_2	1.7 (-2.4, 5.7)	-2.4	0.4	0.5 (-0.5, 1.4)	-1.2 (-5.1, 2.7)
	MR	1.7 (-1.1, 4.5)	-0.0	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.8, 1.3)
All models are restricted	Δ_1	1.7 (-1.1, 4.4)	-3.1	0.2	0.5 (-0.5, 1.4)	-1.2 (-3.7, 1.3)
	Δ_2	1.7 (-1.4, 4.8)	-0.3	0.3	0.5 (-0.5, 1.4)	-1.2 (-4.1, 1.6)
	Δ_3	1.4 (-0.4, 3.2)	-19.7	0.1	0.5 (-0.5, 1.4)	-0.9 (-2.4, 0.6)
	MLE	1.6 (-1.5, 4.7)	-5.6	0.3	0.5 (-0.5, 1.4)	-1.1 (-4.0, 1.7)
	MR	1.7 (-1.2, 4.7)	0.6	0.2	0.5 (-0.5, 1.4)	-1.3 (-4.0, 1.5)

Note: all point estimates and 95% confidence intervals (CI) are scaled by 10^3. Prop bias (%) is the bias calculated as the proportion of the ATE under the saturated model (NP model) taken as the true value.

$H_0 : \Delta = 0$. Similar to the original study, our results indicated a slightly elevated risk of fever among children who received DTaP-IPV-Hib vaccine relative to children who received other DTap-containing comparator vaccines, although the effect was not statistically significant. In addition, there was no evidence of unmeasured confounding as the confidence interval for Δ_{bias} included zero. As expected, under $\mathcal{M}_{\text{nonpar}}$, all methods provided exactly the same point estimate and confidence interval. Under model misspecification, i.e., deviation from the NP model via model restrictions, all methods produced a stable estimate of $\Delta_{\text{confounded}}$, while Δ_{bias} was estimated with larger bias. The MR estimator had generally smaller bias than other methods, which indicated that multiply robust estimation provided protection against model misspecification. A caveat is that in practice if the negative control exposure is rare, the positivity assumption in Assumption 3 may be violated. Sensitivity analysis switching the negative control variables produced similar conclusions, which are presented in Section I of the supplementary material.

6 Final remarks

In this paper, we have developed a general semiparametric framework for causal inference in the presence of unmeasured confounding leveraging a pair of negative control exposure and outcome variables. Our method provides an alternative to more conventional methods such as instrumental variable (IV) methods. Particularly, negative controls are sometimes available when a valid IV may not be, in settings such as air pollution studies (Miao and Tchetgen Tchetgen, 2017), genetic research (Gagnon-Bartsch and Speed, 2012), and observational studies using routinely collected healthcare databases such as electronic health records and claims data (Schuemie et al., 2014). In particular, as majority of the variables in administrative healthcare data are documented by medical codes and thus are naturally categorical, we believe our application study demonstrated the promising role of double negative control for detection and control of confounding bias in observational studies using healthcare databases. Our paper also contributes to the literature of differential confounding misclassification since negative controls can also be viewed as mismeasured versions of the unobserved confounder (Kuroki and Pearl, 2014; Ogburn and VanderWeele, 2012; Miao et al., 2018). Our findings established a theoretical basis for future research on semiparametric estimation with negative control adjustment for continuous unmeasured confounding. Another open problem is the possibility of using modern machine learning for estimation of high dimensional nuisance parameters in the context of multiply robust estimation much in the spirit of Athey and Wager (2017); Chernozhukov et al. (2016); Van der Laan and Rose (2011).

Acknowledgement

We thank the associate editor and two referees for their helpful comments. Research reported in this publication was supported by the National Institutes of Health under award number R01AI104459 (to Eric Tchetgen Tchetgen), Beijing Natural Science Foundation, China under award number Z190001 (to Wang Miao), and Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education, China (to Wang Miao). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix

A.1 Estimation under $\mathcal{M}_1$-$\mathcal{M}_3$

Throughout we use $\dim(v)$ to denote the length of a vector v, such as $\dim(\beta^R)$.

A.1.1 Estimation under $\mathcal{M}_1$

The first class of estimators involves models $f(A, Z \mid X; \alpha^{A,Z})$ and $R(A, X; \beta^R)$ under $\mathcal{M}_1$, with nuisance parameter $\gamma_1 = (\alpha^{A,Z}, \beta^R)$. Specifically, let $\hat{\alpha}^{A,Z}_{\text{mle}}$ denote the MLE of $\alpha^{A,Z}$, and define $f(A \mid Z, X; \hat{\alpha}^{A,Z}_{\text{mle}}) = f(A, Z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})/\sum_a f(A = a, Z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})$ and $f(Z \mid A, X; \hat{\alpha}^{A,Z}_{\text{mle}}) = f(A, Z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})/\sum_z f(A, Z = z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})$. Because $R(A, X; \beta^R)$ does not by itself give rise to a likelihood function, we obtain an estimator $\hat{\beta}^R_{\text{gest}}$ of β^R by solving the following g-estimation type equation (Robins, 1994; Wang and Tchetgen Tchetgen, 2018)

$$\mathbb{P}_n\Big\{ \Big[h_1(A, Z, X) - E[h_1(A, Z, X) \mid A, X; \hat{\alpha}^{A,Z}_{\text{mle}}]\Big] \Big[Y - W \cdot R(A, X; \hat{\beta}^R_{\text{gest}})\Big] \Big\} = 0,$$

where $h_1(A, Z, X)$ is a vector of user-specified $\dim(\beta^R)$ functions of A, Z, and X, and $E[h_1(A, Z, X) \mid A, X; \hat{\alpha}^{A,Z}_{\text{mle}}]$ is evaluated under $f(Z \mid A, X; \hat{\alpha}^{A,Z}_{\text{mle}})$. We then have $\hat{\Delta}_1 = \hat{\Delta}_{\text{confounded,ipw}} - \hat{\Delta}_{\text{bias,gest}}$, where $\hat{\Delta}_{\text{confounded,ipw}} = \mathbb{P}_n \left[\frac{(2A-1)Y}{f(A|Z,X;\hat{\alpha}^{A,Z}_{\text{mle}})}\right]$, $\hat{\Delta}_{\text{bias,gest}} = \mathbb{P}_n \left[E[R(1-A, X; \hat{\beta}^R_{\text{gest}}) \mid Z, X; \hat{\alpha}^{A,Z}_{\text{mle}}]\frac{(2A-1)W}{f(A|Z,X;\hat{\alpha}^{A,Z}_{\text{mle}})}\right]$.

A.1.2 Estimation under $\mathcal{M}_2$

The second class of estimators involves models $f(A, Z \mid X; \alpha^{A,Z})$, $\xi^W_Z(A, X; \beta^{WZ})$, and $\delta^W_A(Z, X; \beta^{WA})$ under $\mathcal{M}_2$, with nuisance parameter $\gamma_2 = (\alpha^{A,Z}, \beta^{WZ}, \beta^{WA})$. Specifically, let $\hat{\beta}^{WZ}_{\text{ipw}}$ and $\hat{\beta}^{WA}_{\text{ipw}}$ solve the following g-estimating equation

$$\mathbb{P}_n\Big\{\Big[h_2(A, Z, X) - E[h_2(A, Z, X) \mid X; \hat{\alpha}^{A,Z}_{\text{mle}}]\Big]\Big[W - \xi^W_Z(A = 0, X; \hat{\beta}^{WZ}_{\text{ipw}})Z - \delta^W_A(Z = 0, X; \hat{\beta}^{WA}_{\text{ipw}})A - \eta^W_{AZ}(X; \hat{\beta}^{WAZ}_{\text{ipw}})AZ\Big]\Big\} = 0$$

where $h_2(A, Z, X)$ is a vector of user-specified functions with dimension $\dim(\beta^{WZ}) + \dim(\beta^{WA}) - \dim(\beta^{WAZ})$, and $E[h_2(A, Z, X) \mid X; \hat{\alpha}^{A,Z}_{\text{mle}}]$ is evaluated under $f(A, Z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})$. Then $\hat{\Delta}_2 = \hat{\Delta}_{\text{confounded,ipw}} - \hat{\Delta}_{\text{bias,ipw}}$, where $\hat{\Delta}_{\text{confounded,ipw}} = \mathbb{P}_n \left[\frac{2A-1}{f(A|Z,X;\hat{\alpha}^{A,Z}_{\text{mle}})}Y\right]$, $\hat{\Delta}_{\text{bias,ipw}} = \mathbb{P}_n\Big[\frac{(2Z-1)Y}{f(Z|A,X;\hat{\alpha}^{A,Z}_{\text{mle}})} \frac{f(1-A|X;\hat{\alpha}^{A,Z}_{\text{mle}})}{f(A|X;\hat{\alpha}^{A,Z}_{\text{mle}})} \frac{E[\delta^W_A(Z,X;\hat{\beta}^{WA}_{\text{ipw}})|1-A,X;\hat{\alpha}^{A,Z}_{\text{mle}}]}{\xi^W_Z(A,X;\hat{\beta}^{WZ}_{\text{ipw}})}\Big]$.

A.1.3 Estimation under $\mathcal{M}_3$

The third class of estimators involves models $E[W \mid Z, A, X; \beta^W]$, $E[Y \mid Z = 0, A, X; \beta^Y]$, and $R(A, X; \beta^R)$ under $\mathcal{M}_3$, with nuisance parameter $\gamma_3 = (\beta^W, \beta^Y, \beta^R)$. Specifically, let $\hat{\beta}^W_{\text{mle}} = (\hat{\beta}^{W0}_{\text{mle}}, \hat{\beta}^{WZ}_{\text{mle}}, \hat{\beta}^{WA}_{\text{mle}})$ denote the MLE of β^W, and $\hat{\beta}^Y_{\text{mle}}$ denote the restricted MLE of β^Y, where the latter is obtained by maximizing the likelihood under the working model $E[Y \mid Z = 0, A, X; \beta^Y]$ restricted to the subsample with $Z = 0$. Let $\hat{\beta}^R_{\text{or}}$ solve the following estimating equation

$$\mathbb{P}_n\Big[h_3(A, Z, X)\Big(Y - E[Y \mid Z = 0, A, X; \hat{\beta}^Y_{\text{mle}}] - R(A, X; \hat{\beta}^R_{\text{or}})(W - E[W \mid Z = 0, A, X; \hat{\beta}^W_{\text{mle}}])\Big)\Big] = 0,$$

where $h_3(A, Z, X)$ is a nonzero vector function of dimension $\dim(\beta^R)$. We obtain $E[Y \mid Z, A, X; \hat{\beta}^Y_{\text{mle}}, \hat{\beta}^W_{\text{mle}}; \hat{\beta}^R_{\text{or}}]$ by Eq. (9) using $E[Y \mid Z = 0, A, X; \hat{\beta}^Y_{\text{mle}}]$, $\xi^W_Z(A, X; \hat{\beta}^W_{\text{mle}})$, and $R(A, X; \hat{\beta}^R_{\text{or}})$. Combining the above estimators, we have $\hat{\Delta}_3 = \hat{\Delta}_{\text{confounded,or}} - \hat{\Delta}_{\text{bias,or}}$, where $\hat{\Delta}_{\text{confounded,or}} = \mathbb{P}_n\Big[E[Y \mid A = 1, Z, X; \hat{\beta}^Y_{\text{mle}}, \hat{\beta}^W_{\text{mle}}; \hat{\beta}^R_{\text{or}}] - E[Y \mid A = 0, Z, X; \hat{\beta}^Y_{\text{mle}}, \hat{\beta}^W_{\text{mle}}; \hat{\beta}^R_{\text{or}}]\Big]$ and $\hat{\Delta}_{\text{bias,or}} = \mathbb{P}_n\Big[R(1-A, X; \hat{\beta}^R_{\text{or}})\delta^W_A(Z, X; \hat{\beta}^W_{\text{mle}})\Big]$.

A.2 Estimated working models for the multiply robust estimator

Following the variation independent parameterization detailed in (7)-(10), we specify the estimated working models by plugging in the corresponding components in θ as follows: $f(A \mid Z, X; \hat{\theta}) = f(A, Z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}}) / \sum_a f(A = a, Z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})$, $f(A \mid X; \hat{\theta}) = \sum_z f(A, Z = z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})$, $f(Z \mid A, X; \hat{\theta}) = f(A, Z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}}) / \sum_z f(A, Z = z \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})$, $E[Y \mid A = 0, Z, X; \hat{\theta}] = E[Y \mid Z = 0, A, X; \hat{\beta}^Y_{\text{mle}}] + R(A, X; \hat{\beta}^R_{\text{dr}})\xi^W_Z(A, X; \hat{\beta}^{WA}_{\text{dr}})Z$, $E[Y \mid Z, A, X; \hat{\theta}] = E[Y \mid Z = 0, A, X; \hat{\beta}^Y_{\text{mle}}] + R(A, X; \hat{\beta}^R_{\text{dr}})\xi^W_Z(A, X; \hat{\beta}^{WZ}_{\text{dr}})$, $E[W \mid A, Z, X; \hat{\theta}] = E[W \mid A = 0, Z = 0, X; \beta^{W0}_{\text{mle}}] + \xi^W_Z(A = 0, X; \beta^{WZ}_{\text{dr}})Z + \delta^W_A(Z = 0, X; \beta^{WA}_{\text{dr}})A + \eta^W_{AZ}(X; \beta^{WAZ}_{\text{dr}})AZ$, $E[R(1 - A, X) \mid Z, X; \hat{\theta}] = \sum_a R(1 - a, X; \hat{\beta}^R_{\text{dr}})f(A = a \mid Z, X; \hat{\alpha}^{A,Z}_{\text{mle}})$, and $E[\delta^W_A(Z, X) \mid 1 - A, X; \hat{\theta}] = \sum_z \delta^W_A(z, X; \hat{\beta}^{WA}_{\text{dr}})f(Z = z \mid 1 - A, X; \hat{\alpha}^{A,Z}_{\text{mle}})$. In addition, to simplify notation, we let $R(A, X; \hat{\theta}) = R(A, X; \hat{\beta}^R_{\text{dr}})$, $\delta^W_A(Z, X; \hat{\theta}) = \delta^W_A(Z, X; \hat{\beta}^{WA}_{\text{dr}})$, and $\xi^W_Z(A, X; \hat{\theta}) = \xi^W_Z(A, X; \hat{\beta}^{WZ}_{\text{dr}})$.

References

Athey, S. and Wager, S. (2017). Efficient policy learning. *arXiv preprint arXiv:1702.02896*.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bickel, P. J., Klaassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and adaptive inference in semiparametric models*, volume 2. Baltimore: Johns Hopkins University Press.

Cheng, G., Huang, J. Z., et al. (2010). Bootstrap consistency for general semiparametric M-estimation. *The Annals of Statistics*, 38(5):2884–2915.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Choi, Y., Taylor, J., Tibshirani, R., et al. (2017). Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *Annals of Statistics*, 45(6):2590–2617.

Davey Smith, G. (2008). Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings? *Basic & Clinical Pharmacology & Toxicology*, 102(2):245–256.

Davey Smith, G. (2012). Negative control exposures in epidemiologic studies. Comments on "Negative controls: a tool for detecting confounding and bias in observational studies". *Epidemiology*, 23(2):350–351.

Flanders, W. D., Klein, M., Darrow, L. A., Strickland, M. J., Sarnat, S. E., Sarnat, J. A., Waller, L. A., Winquist, A., and Tolbert, P. E. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*, 22(1):59.

Flanders, W. D., Strickland, M. J., and Klein, M. (2017). A new method for partial correction of residual confounding in time-series and other observational studies. *American Journal of Epidemiology*, 185(10):941–949.

Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552.

Glass, D. J. (2014). *Experimental design for biologists*. Cold Spring Harbor Laboratory Press.

Gómez, P., Lacruz, B., and Pruneda, R. E. (2008). Generalized inverse computation based on an orthogonal decomposition methodology. In *Advances in Mathematical and Statistical Modeling*, pages 305–315. Springer.

Jackson, L. A., Jackson, M. L., Nelson, J. C., Neuzil, K. M., and Weiss, N. S. (2005). Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International Journal of Epidemiology*, 35(2):337–344.

James, M. (1978). The generalised inverse. *The Mathematical Gazette*, 62(420):109–114.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Lipsitch, M., Tchetgen Tchetgen, E. J., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383.

Lipsitch, M., Tchetgen Tchetgen, E. J., and Cohen, T. (2012). Negative control exposures in epidemiologic studies (author reply). *Epidemiology*, 23(2):351–352.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Miao, W., Shi, X., and Tchetgen, E. T. (2019). A confounding bridge approach for double negative control inference on causal effects. In progress, a prior version can be found at `https://arxiv.org/abs/1808.04945`.

Miao, W. and Tchetgen Tchetgen, E. J. (2017). Invited commentary: Bias attenuation and identification of causal effects with multiple negative controls. *American Journal of Epidemiology*, 185(10):950–953.

Nelson, J. C., Yu, O., Dominguez-Islas, C. P., Cook, A. J., Peterson, D., Greene, S. K., Yih, W. K., Daley, M. F., Jacobsen, S. J., Klein, N. P., et al. (2013). Adapting group sequential methods to observational postlicensure vaccine safety surveillance: results of a pentavalent combination DTaP-IPV-Hib vaccine safety study. *American Journal of Epidemiology*, 177(2):131–141.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

Ogburn, E. L. and VanderWeele, T. J. (2012). On the nondifferential misclassification of a binary confounder. *Epidemiology*, 23(3):433.

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*, page 128.

Robins, J. M. (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. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and Methods*, 23(8):2379–2412.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, 1999:6–10.

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

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.

Rosenbaum, P. R. (1989). The role of known effects in observational studies. *Biometrics*, 45(2):557–569.

Rosenbaum, P. R. (1992). Detecting bias with confidence in observational studies. *Biometrika*, 79(2):367–374.

Rotnitzky, A., Robins, J., and Babino, L. (2017). On the multiply robust estimation of the mean of the g-functional. *arXiv preprint arXiv:1705.08582*.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., and Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in Medicine*, 33(2):209–218.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen Tchetgen, E. J. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science*, 31(3):348.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Tchetgen Tchetgen, E. J. (2013). The control outcome calibration approach for causal inference with unobserved confounding. *American Journal of Epidemiology*, 179(5):633–640.

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.

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

Van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

Vansteelandt, S., VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Robins, J. M. (2008). Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, 103(484):1693–1704.

Wang, J., Zhao, Q., Hastie, T., Owen, A. B., et al. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of Statistics*, 45(5):1863–1894.

Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550.

Weiss, N. S. (2002). Can the "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology*, 13(1):6–8.