Robust inference on indirect causal effects

Isabel R. Fulcher
Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA

Ilya Shpitser
Department of Computer Science, Johns Hopkins University, Baltimore, USA

Stella Marealle
D-tree International, Zanzibar, Tanzania

Eric J. Tchetgen Tchetgen
Wharton Statistics Department, University of Pennsylvania, Philadelphia, USA

Abstract

Standard methods for inference about direct and indirect effects require stringent no unmeasured confounding assumptions which often fail to hold in practice, particularly in observational studies. The goal of this paper is to introduce a new form of indirect effect, the population intervention indirect effect (PIIE), that can be nonparametrically identified in the presence of an unmeasured common cause of exposure and outcome. This new type of indirect effect captures the extent to which the effect of exposure is mediated by an intermediate variable under an intervention which fixes the component of exposure directly influencing the outcome at its observed value. The PIIE is in fact the indirect component of the population intervention effect, introduced by Hubbard and Van der Laan (2008). Interestingly, our identification criterion relaxes Judea Pearl’s front-door criterion as it does not require no direct effect of exposure not mediated by the intermediate variable. For inference, we develop both parametric and semiparametric methods, including a doubly robust semiparametric locally efficient estimator, that perform very well in simulation studies. Finally, the proposed methods are used to measure the effectiveness of monetary saving recommendations among women enrolled in a maternal health program in Tanzania.

Key words: double robust, indirect effects, front-door criterion, mediation analysis, population intervention effect
1 Introduction

Recent advances in causal inference have formalized causal and statistical conditions for identification and estimation of direct and indirect effects. Natural (pure) direct and indirect effects have emerged as the most common form of causal effects in modern mediation analysis. Identification of these effects typically relies on an assumption that there is no unmeasured confounding of the exposure-outcome, exposure-mediator, and mediator-outcome relationships (Pearl, 2001; Avin et al., 2005). The assumption of no unmeasured confounding of the effects of exposure is typically guaranteed by design in randomized experiments. In settings where a randomized experiment is impractical or unethical, observational data may instead be used, in which case unmeasured confounding cannot be ruled out.

In this paper, we introduce a new form of indirect effect that has a causal interpretation as it can be expressed as a contrast of potential outcomes, much like the natural indirect effect (NIE). However, unlike the NIE, the newly defined indirect effect can be identified even when there is an unmeasured common cause of exposure and outcome variables, provided it is not also a cause of the mediator. Our newly defined indirect effect is particularly attractive for settings where the exposure is known or suspected to be harmful, so that one may not want to conceive of an intervention that forces a person to be exposed. For instance, pharmacoepidemiologists are often concerned that medications prescribed for serious health conditions, such as a disease, may have unintended adverse health consequences (Tchetgen Tchetgen and Phiri, 2017). Thus, it is of interest to understand the effect of having a certain disease on undesirable health outcomes through a pathway mediated by use of prescribed medication. For example, in HIV research, studies have shown that there is an increased risk in adverse birth outcomes for HIV positive women who use antiretroviral treatment during pregnancy compared to those who do not (Cotter et al., 2006; Martin and Taylor, 2007; Machado et al., 2008). However, antiretroviral treatment is the standard of care for pregnant women because it slows disease progression, improves health of infected women, and prevents mother-to-child HIV transmission. In a separate example, a harmful association has previously been established between the use of anti-epileptic drugs during pregnancy and adverse birth outcomes (Holmes et al., 2001). In both of these examples, it is primarily of interest to quantify the extent to which the causal effects of maternal disease status (HIV status, epilepsy status) on birth outcomes operates through maternal medication use during pregnancy. In each of these examples, one could in principle use the NIE to address the query of interest. However, two complications arise.

First, interpretation of the NIE requires conceiving of an intervention that would force a pregnant woman to become HIV positive or to acquire epilepsy. Such intervention is unlikely
to be of scientific relevance as it would realistically not be considered in practice. Second, because such observational studies are typically based on retrospective clinical chart abstractions that seldom include information on behavioral risk factors for HIV and birth outcomes such as smoking and alcohol use, unmeasured confounding is of concern, particularly between exposure and outcome variables. For this reason, existing mediation methods cannot be used without making additional assumptions. Tchetgen Tchetgen and Phiri (2017) identified the NIE among HIV positive mothers, but they relied on an additional exclusion restriction that the mediator potential outcome under no exposure is degenerate; specifically, they assumed that HIV negative mothers are unexposed to ART with probability 1. This assumption may apply if HIV negative women never take antiretroviral therapy. However, this assumption may be violated in the advent of pre-exposure prophylaxis therapy whereby HIV negative women may be taking ARTs to reduce the chance of HIV infection by a partner known to be HIV seropositive. Additionally, the exclusion restriction assumption is not necessarily tenable for epilepsy because the anti-epileptic drugs can be prescribed to treat psychiatric disorders and chronic pain (Ettinger and Argoff, 2007). In this paper, we do not rely on the exclusion restriction of Tchetgen Tchetgen and Phiri (2017) to identify the PIIE of a confounded exposure.

The population intervention indirect effect (PIIE) can be interpreted as the contrast between the observed outcome mean for the population and the population had contrary to fact the mediator taken the value it would have in the absence of exposure. In the HIV example, the PIIE is the difference in the observed prevalence of adverse birth outcomes compared to the prevalence of adverse birth outcomes under a hypothetical intervention which sets maternal ART exposure to what it would be in the absence of HIV infection. Although medication-mediated effects provide a compelling motivation for the PIIE, this estimand may be of interest in a variety of other settings when both the exclusion restriction does not hold and unmeasured confounding of the exposure-outcome relation cannot be ruled out with certainty. One such example is given later in the paper where we investigate the indirect effect of a woman’s pregnancy risk on monetary savings for delivery mediated by the amount she is recommended to save by a community health worker.

The PIIE and its identifying formula connects to prior literature in two important ways. First, the PIIE may be viewed as the indirect component of the population intervention effect (PIE) defined by Hubbard and Van der Laan (2008). The PIE of an exposure is a form of total effect relating the mean of an outcome in the population to that in the population had contrary to fact no one been exposed to the exposure (Hubbard and Van der Laan, 2008). In many public health settings, the PIE may be more relevant than the population average causal effect, particularly when evaluating the potential impact of programs that eliminate
a harmful exposure, in which case a hypothetical intervention were everyone is exposed is of no scientific interest. Note that while the PIIE is identified in presence of confounding of exposure, both the PIE and the direct component of the PIE, which we call the population intervention direct effect (PIDE), cannot generally be identified under the same conditions (Hubbard and Van der Laan, 2008). Second, the identifying formula we obtain for a portion of the PIIE matches Judea Pearl’s celebrated front-door formula of the indirect effect of a confounded exposure (Pearl, 2009). However, whereas identification of indirect effects with Pearl’s front-door criterion requires a key assumption of no direct effect of the exposure on the outcome not through the mediator in view, our generalized front-door criterion allows for presence of such direct effects. We emphasize that while the front-door criterion has long been established, the proposed generalized front-door criterion is entirely new to the literature. In addition to new identification results, we also develop both parametric and semiparametric theory for inference about the PIIE. This latter contribution of the paper may be of independent interest as to the best of our knowledge it also delivers the first doubly robust estimator of Pearl’s front-door formula in the literature.

The rest of the paper is organized as follows, in section 2, we discuss nonparametric identification of the PIIE. In section 3, we derive both parametric and semiparametric estimators, including a doubly robust semiparametric locally efficient estimator for the PIIE. In section 4, the performance of these estimators is evaluated in a range of settings in extensive simulation studies. In section 5, the proposed methods are used to measure the effectiveness of monetary savings recommendations for delivery among pregnant women enrolled in a maternal health program in Zanzibar, Tanzania.

2 Nonparametric Identification

In the following, let $Z(a)$ denote the counterfactual mediator variable had the exposure taken value $a$ and $Y(a) = Y(a, Z(a))$ denote the counterfactual outcome had exposure possibly contrary to the fact taken value $a$. We will also consider the counterfactual outcome $Y(A, Z(a^*)) = Y(Z(a^*))$ had exposure taken its natural level and the mediator variable taken the value it would have under $a^*$. Note that when $a^* = 0$, $Y(Z(0))$ is the counterfactual outcome had exposure taken its natural level and the mediator variable taken the value it would have under no exposure. Additionally, let $C$ be a set of observed pre-exposure covariates known to confound $A-Z$, $A-Y$ and $Z-Y$ associations. Throughout $Z$ can be vector valued.

We first consider the standard decomposition of the average total effect (ATE). For
exposure levels \(a\) and \(a^{*}\),

\[
ATE(a,a^{*}) = E[Y(a,Z(a)) - Y(a^{*},Z(a^{*}))]
\]

\[
= E[Y(a,Z(a)) - Y(a,Z(a^{*}))] + E[Y(a,Z(a^{*})) - Y(a^{*},Z(a^{*}))]
\]

The natural indirect effect is the difference between the potential outcome under exposure value \(a\) and the potential outcome had exposure taken value \(a\) but the mediator variable had taken the value it would have under \(a^{*}\);

\[
NIE(a,a^{*}) = E[Y(a,Z(a)) - Y(a,Z(a^{*}))]
\]

The natural direct effect is therefore given by \(ATE(a,a^{*}) - NIE(a,a^{*})\). The NIE and NDE are well-known to be identified under the following conditions (Pearl, 2012; Imai et al., 2010):

M1. Consistency assumptions: (1) If \(A = a\), then \(Z(a) = Z\) w.p.1,

(2) If \(A = a\), then \(Y(a) = Y\) w.p.1,

(3) If \(A = a\) and \(Z = z\), then \(Y(a,z) = Y\) w.p.1

M2. \(Z(a^{*}) \perp A \mid C = c\) \(\forall a^{*},c\)

M3. \(Y(a,z) \perp Z(a^{*}) \mid A = a, C = c\) \(\forall z,a,a^{*},c\)

M4. \(Y(a,z) \perp Z \mid A = a, C = c\) \(\forall z,a,c\)

M5. \(Y(a,z) \perp A \mid C = c\) \(\forall z,a,c\)

M1 states the observed outcome is equal to the counterfactual outcome corresponding to the observed treatment. The remaining assumptions essentially state that there is no unmeasured confounding of the exposure and the mediator variable (M2), the mediator variable and the outcome (M3, M4), and the exposure and the outcome (M5). These assumptions could equivalently be formulated under a Nonparametric Structural Equation Model with Independent Errors (NPSEM-IE) interpretation of the diagram in Figure 1a (Pearl, 2009). Under M1-5,

\[
E[Y(a)] = \sum_{c} E(Y \mid A = a, Z = z, C = c)Pr(A = a \mid C = c)Pr(C = c) \quad (1)
\]

\[
E[Y(a,Z(a^{*}))] = \sum_{c,z} E(Y \mid A = a, Z = z, C = c)Pr(Z = z \mid A = a^{*}, C = c)Pr(C = c) \quad (2)
\]
The NIE fails to be nonparametrically identified if any of assumptions M1-5 fail to hold without an additional assumption (Shpitser, 2013).

We will now formally define the population intervention indirect effect, a novel measure of indirect effect corresponding to the effect of an intervention which changes the mediator from its natural value (i.e. its observed value) to the value it would have had under exposure value $a^*$,

$$PIIE(a^*) = E[Y(A, Z(A)) - Y(A, Z(a^*))] = E[Y - Y(Z(a^*))]$$

(3)

This is indeed an indirect effect as it would only be non-null if changing the exposure from its natural value to $a^*$ results in a change in the value of the mediator which in turn results in a change in the value of the outcome. That is, the PIIE captures an effect propagated along the $A \rightarrow Z \rightarrow Y$ pathway only, and would be null for a given person either if $A$ has no effect on $Z$ or $Z$ has no effect on $Y$ for the person. Compared to the NIE, the PIIE only requires intervention on the exposure level of the mediator in the second term and does not require intervention on the exposure level for the potential outcomes for $Y$.

The first term of the PIIE, $E(Y)$, is nonparametrically identified; however, the second term requires identification conditions. Identification conditions for the PIIE are less stringent than the NIE as seen by comparing Figure 1a and 1c under a NPSEM-IE interpretation of the diagrams (Pearl, 2009). In fact, the following result states that assumption M5 is no longer needed.

Lemma 1 Under assumptions M1-4, the population intervention indirect effect is given by,

$$PIIE(a^*) = E[Y] - E[Y(Z(a^*))] = E[Y] - \Psi$$

where

$$\Psi = \sum_{z,c} Pr(Z = z \mid A = a^*, C = c) \times \sum_a E(Y \mid A = a, Z = z, C = c) Pr(A = a \mid C = c) Pr(C = c)$$

(4)

Further, equation (4) implies nonparametric identification in the sense that it does not restrict the observed data distribution. The proof for this lemma can be found in the appendix.

Interestingly, $\Psi$ is closely connected to Judea Pearl’s front-door criterion. Pearl’s front-door criterion provides conditions for identification of the indirect effect in the presence of unmeasured confounding of the exposure-outcome relation. The criterion requires: (1) $Z$...
intercepts all directed paths from the exposure $A$ to the outcome $Y$ so that the indirect
effect equals the total effect of $A$ on $Y$, (2) there is no unblocked back-door path from $A$ to $Z$, and (3) all back-door paths from $Z$ to $Y$ are blocked by $A$ (Pearl, 2009). More formally, assume that M1-4 and the additional assumption hold,

$$ F1. Y(a, z) = Y(a^*, z) = Y(z) \forall a, a^*, z $$

F1 crucially states that $Z$ fully mediates the effect of $A$ on $Y$. In other words, mediator variable(s) $Z$ intercepts all directed paths from the exposure to the outcome. Figure 1b encodes one possible graph that satisfies the front-door criterion under a Finest Fully Randomized Causally Interpretable Structured Tree Graph, a submodel of the NPSEM-IE, interpretation of the causal diagram (Pearl, 2009, 2012; Robins, 1986).

When F1 holds, the cross-world term $E(Y(Z(a^*))) = E(Y(a^*))$. The identifying formula for the latter term is known as Pearl’s front-door functional and matches equation (4) (Pearl, 2009). Thus, under the front-door criterion M1-4 and F1, the population intervention indirect effect can be expressed as,

$$ PIIE(a^*) = E[Y] - E[Y(a^*)] \quad (5) $$

Interestingly, The right hand side of equation (5) corresponds to the PIE of Hubbard and Van der Laan (2008) and defines a total effect of $A$ irrespective of whether F1 holds or not. Without F1, the population intervention direct effect (PIDE) is given by $PIE(a^*) - PIIE(a^*)$. That is,

$$ PIDE(a^*) = E[Y(Z(a^*)) - Y(a^*)] $$

The PIDE is a novel measure of direct effect corresponding to the effect of an intervention which changes the exposure from its natural level to the value under intervention $a^*$, while keeping the mediator variable at the value it would have under intervention $a^*$. This is indeed a direct effect as it would only be non-null if changing the exposure from its natural value to $a^*$, while preventing the mediator variable to change, results in a change in the value of the outcome. That is, the PIDE captures an effect along the $A \rightarrow Y$ pathway only. Under F1, the PIDE is null, and the PIE reduces to equation (5).

The identifying conditions for the PIIE can be thought of as a generalization of Pearl’s front-door criterion as F1 need not hold, thereby allowing a direct effect of the exposure $A$ on the outcome $Y$, not through the mediator variable(s) $Z$ (i.e. the PIDE may or may not be null). Importantly, while the PIIE is nonparametrically identified under M1-4, the PIE and the PIDE are not identified. In the event that M5 also holds, the PIE and PIDE are
both nonparametrically identified along with the NIE and PIIE.

We briefly note that in the special case of binary $A$, the PIE can be written as the effect of treatment on the treated (ETT) scaled by prevalence of treated persons,

\[ PIE(0) = \frac{E(Y(1) - Y(0)|A = 1)}{Pr(A = 1)} \]

See proof in appendix A2.5. Thus, the PIIE and PIDE can respectively be written as the indirect and direct components of the ETT simply upon rescaling by the prevalence of treated persons. This decomposition of the ETT offers an alternative to that of Vansteelandt and VanderWeele (2012). Further discussion can be found in appendix section A2.6.

### 3 Estimation and Inference

#### 3.1 Parametric estimation

We have considered identification under a nonparametric model for the observed data distribution. Estimation of formula (4) clearly requires estimation of the mean of $Y|A, Z, C$ and the densities for $Z|A, C$, $A|C$, and $C$. In principle, one may wish to estimate this quantities nonparametrically; however, as will typically be the case in practice, the observed set of covariates $C$ may have two or more components that are continuous, so that the curse of dimensionality would rule out the use of nonparametric estimators such as kernel smoothing or series estimation. Thus, we propose four estimators for the population intervention indirect effect that impose parametric models for different parts of the observed data likelihood, allowing other parts to remain unrestricted. Under this setting, each estimator will be consistent and asymptotically normal (CAN) under the assumed semiparametric model. We also propose a doubly robust estimator which is CAN under a semiparametric union model thereby allowing for robustness to partial model misspecification.

We only discuss estimation for the second term in the PIIE contrast, $\Psi$, as the first term $E(Y)$ can be consistently estimated nonparametrically by the empirical mean of $Y$. Let $Pr(y|a, z, c; \theta)$ denote a model for the density of $Y|A, Z, C$ evaluated at $y, a, z, c$ and indexed by $\theta$. Likewise, let $Pr(z|a, c; \beta)$and $Pr(a|c; \alpha)$ denote models for $Z|A, \ C$ and $A|C$ evaluated at $z, a, c$ and $a, c$ respectively with corresponding parameters $\beta$ and $\alpha$. These models could in principle be made as flexible as allowed by sample size, to simplify exposition, we will focus on simple parametric models. The first of the four estimators is the maximum likelihood estimator (MLE), $\hat{\Psi}_{mle}$, under a model that specifies parametric models for $A$, $Z$, and $Y$, and a nonparametric model for the distribution of $C$ estimated by its empirical distribution.
Figure 1: Causal diagrams with indirect effects as dashed lines. The following indirect effects are identified in each diagram under a Nonparametric Structural Equation Model with Independent Errors (Pearl, 2009) interpretation of the diagram: (a) natural indirect effect and population intervention indirect effect, (b) natural indirect effect (equal to the total effect) and population intervention indirect effect (equal to the population intervention effect), and (c) population intervention indirect effect. Further, the indirect effects in (b) are identified under a Finest Fully Randomized Causally Interpretable Structured Tree Graph (Robins, 1986).
The MLE is obtained by the plug-in principle (Casella and Berger, 2002):

\[
\hat{\Psi}_{mle} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{z} Pr(Z = z \mid A = a^*, C_i = c_i; \hat{\beta}) \times \right. \\
\left. \sum_{a} E(Y \mid A = a, Z = z, C_i = c_i; \hat{\theta}) Pr(A = a \mid C_i = c_i; \hat{\alpha}) \right\}
\]

where \( \hat{\theta}, \hat{\beta}, \) and \( \hat{\alpha} \) are the MLEs of \( \theta, \beta, \) and \( \alpha. \) This estimator is only consistent under correct specification of the three required models, which we define as \( M_{y,z,a}. \) An example of the MLE when \( Y \) and \( Z \) are continuous is given in the appendix.

3.2 Semiparametric estimation

Next, we consider two semiparametric estimators for \( \Psi. \) The first is under model \( M_z \) which posits a parametric model for the law of \( Z \mid A, C \) but allows the densities of \( Y \mid A, Z, C, A \mid C, \) and \( C \) to remain unrestricted. The second is under model \( M_{y,a} \) which instead posits a model for the outcome mean of \( Y \mid Z, A, C \) and the density of \( A \mid C \) but allows the densities of \( Z \mid A, C \) and \( C \) to be unrestricted.

\[
\hat{\Psi}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i f(Z_i \mid a^*, C_i; \hat{\beta})}{f(Z_i \mid A_i, C_i; \hat{\beta})} \\
\hat{\Psi}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a^*)}{f(A_i \mid C_i; \hat{\alpha})} E \left( E \{ Y_i \mid A_i, Z_i, C_i; \hat{\theta} \} \mid C_i; \hat{\alpha} \right)
\]

Define the following positivity assumptions,

P1. There exists \( m_1 > 0 \) such that \( f(Z \mid A, C) > m_1 \) almost surely

P2. There exists \( m_2 > 0 \) such that \( f(A \mid C) > m_2 \) almost surely

Lemma 2 Under standard regularity conditions and P1, the estimator \( \hat{\Psi}_1 \) is consistent and asymptotically normal under model \( M_z. \)

Lemma 3 Under standard regularity conditions and P2, the estimator \( \hat{\Psi}_2 \) is consistent and asymptotically normal under model \( M_{y,a}. \)

The estimator \( \hat{\Psi}_1 \) will generally fail to be consistent if the density for \( Z \mid A, C \) is incorrectly specified even if the rest of the likelihood is correctly specified. Likewise, the estimator \( \hat{\Psi}_2 \)
will also generally fail to be consistent if either the mean model for $Y|A, Z, C$ or the density of $A|C$ is incorrectly specified. In order to motivate our doubly robust estimator, the following result gives the efficient influence function for $\Psi$ in the nonparametric model $\mathcal{M}_{np}$, which does not place any model restriction on the observed data distribution. The following results are entirely novel and have previously not appeared in the literature.

**Theorem 1** The efficient influence function of $\Psi$ in $\mathcal{M}_{np}$ is:

$$
\varphi_{eff}(Y, Z, A, C) = (Y - E(Y | A, Z, C)) \frac{f(Z | a^*, C)}{f(Z | A, C)} + \frac{I(A = a^*)}{f(A | C)} \left( \sum_a E[Y | a, Z, C] f(a | C) - \sum_{a, \tilde{z}} E(Y | a, \tilde{z}, C) f(\tilde{z} | A, C) f(a | C) \right)
$$

$$
+ \sum_z E[Y | A, z, C] f(z | a^*, C) - \Psi
$$

and the semiparametric efficiency bound of $\Psi$ in $\mathcal{M}_{np}$ is given by $\text{var}\{\varphi_{eff}\}$.

The proof for this theorem can be found in the appendix. An implication of this result is that for any regular and asymptotically linear (RAL) estimator $\hat{\Psi}$ in model $\mathcal{M}_{np}$ it must be that $\sqrt{n}(\hat{\Psi} - \Psi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{eff}(Y_i, Z_i, A_i, C_i) + o_p(1)$. In other words, all RAL estimators in this model are asymptotically equivalent and attain the semiparametric efficiency bound of $\Psi$ in $\mathcal{M}_{np}$ (Bickel et al., 1998). The result motivates the following estimator of $\Psi$, which we formally establish to be doubly robust.

$$
\hat{\Psi}_{dr} = \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i - E(Y | A_i, Z_i, C_i; \hat{\theta}) \right] \frac{f(Z | a^*, C_i; \hat{\beta})}{f(Z | A_i, C_i; \hat{\beta})} + \frac{I(A_i = a^*)}{Pr(A_i = a^* | C_i; \hat{\alpha})} \left( \sum_a E[Y | a, Z_i, C_i; \hat{\theta}] f(a | C_i; \hat{\alpha}) - \sum_{a, \tilde{z}} E(Y | a, \tilde{z}, C_i; \hat{\theta}) f(\tilde{z} | A_i, C_i; \hat{\beta}) f(a | C_i; \hat{\alpha}) \right)
$$

$$
+ \sum_z E[Y | A_i, z, C_i; \hat{\theta}] f(z | a^*, C_i; \hat{\beta})
$$

**Theorem 2** Under standard regularity conditions and the positivity assumptions given by $P1$ and $P2$, the estimator $\hat{\Psi}_{dr}$ is consistent and asymptotically normal provided that one
of the following holds: (1) the model for the mean $E(Y|A,C)$ and the exposure density $f(A|C)$ are both correctly specified; or (2) The model for the mediator density $f(Z|A,C)$ is correctly specified. Also, $\hat{Ψ}_{dr}$ attains the semiparametric efficiency bound for the union model $M_{union} = M_{y,a} \cup M_z$, and therefore for the nonparametric model $M_{np}$ at the intersection submodel where all models are correctly specified.

The estimator $\hat{Ψ}_{dr}$ offers two genuine opportunities to consistently estimate $Ψ$, and, thus, the PIIE. This is clearly an improvement over the other estimators $\hat{Ψ}_{mle}$, $\hat{Ψ}_1$ and $\hat{Ψ}_2$, which are only guaranteed to be consistent under more stringent parametric restrictions. For inference on $Ψ$, we provide a consistent estimator of the asymptotic variance for the proposed estimators in the appendix section A2.4. Wald-type confidence intervals for $Ψ$ can then be based on $\hat{Ψ}_{mle}$, $\hat{Ψ}_1$, $\hat{Ψ}_2$, or $\hat{Ψ}_{dr}$ and the corresponding standard error estimator.

An important advantage of the doubly-robust estimator is that it can easily accommodate modern machine learning for estimation of high dimensional nuisance parameters, such as $E(Y|A,Z,C)$ or $f(Z|A,C)$ (Van der Laan and Rose, 2011; Newey and Robins, 2017; Chernozhukov et al., 2017). Although, investigators should exercise caution when implementing these more flexible methods, particularly if nonparametric methods are used to estimate nuisance parameters. This is because such methods typically cannot attain root-n convergence rates, although the doubly robust estimator would in principle provide valid root-n inferences about $Ψ$ provided that estimators of nuisance parameters have a convergence rate faster than $n^{-1/4}$. A major challenge with using complex machine learning methods such as Random forests arises if the corresponding estimator of the nuisance function (say $f(A|C)$) fails to be consistent at rate $n^{1/4}$ even if other nuisance function (say $f(Z|A,C)$) is estimated at rate root-n, in such case, it is not entirely clear what the asymptotic distribution is for $\hat{Ψ}_{dr}$.

4 Simulation Study

4.1 Data generating mechanism

We now report extensive simulation studies which aim to illustrate: (i) robustness of PIIE to exposure-outcome unmeasured confounding (ii) robustness properties to model misspecification of our various semiparametric estimators. The data generating mechanism for sim-
Simulations was as followed:

\[ C_1 \sim \text{Ber}(0.6) \]
\[ C_2 | C_1 \sim \text{Ber}(\expit(1 + 0.5c_1)) \]
\[ C_3 \sim \text{Ber}(0.3) \]
\[ A | C_1, C_2, C_3 \sim \text{Ber}(\expit(0.5 + 0.2c_1 + 0.4c_2 + 0.5c_1c_2 + 0.2c_3)) \]
\[ Z | A, C_1, C_2 \sim N(1 + a - 2c_1 + 2c_2 + 8c_1c_2, 4) \]
\[ Y | A, Z, C_1, C_2, C_3 \sim N(1 + 2a + 2z - 8az + 3c_1 + c_2 + c_1c_2 + c_3, 1) \]

Therefore, \( C_1, C_2, \) and \( C_3 \) confound the \( A - Y \) association while only \( C_1 \) and \( C_2 \) confound the \( A - M \) and \( M - Y \) associations. Simulations were performed 10,000 times with a sample size of 1,000. We evaluated the performance of the proposed estimators under the following settings,

\[ \mathcal{M}_{y,z,a} : \quad \hat{E}(Y \mid a, z, c_1, c_2, c_3), \quad \hat{f}(Z \mid a, c_1, c_2), \quad \hat{f}(A \mid c_1, c_2, c_3) \]
\[ \mathcal{M}'_{y,z,a} : \quad \tilde{E}(Y \mid a, z, c_1, c_2) \quad (c_3 \text{ left out}), \quad \tilde{f}(A \mid c_1, c_2) \quad (c_3 \text{ left out}), \quad \hat{f}(Z \mid a, c_1, c_2) \]
\[ \mathcal{M}_z : \quad \tilde{E}(Y \mid a, z, c_1, c_2, c_3) \quad (az \text{ left out}), \quad \tilde{f}(A \mid c_1, c_3) \quad (c_2 \text{ and } c_1c_2 \text{ left out}), \quad \hat{f}(Z \mid a, c_1, c_2) \]
\[ \mathcal{M}_{y,a} : \quad \hat{E}(Y \mid a, z, c_1, c_2, c_3), \quad \hat{f}(A \mid c_1, c_3), \quad \tilde{f}(Z \mid a, c_1) \quad (c_2 \text{ and } c_1c_2 \text{ left out}) \]

where \( \hat{\cdot} \) denotes that the model is correctly specified and \( \sim \) and \( - \) denote the model is misspecified. Note that we did not specify a model for \( A \) for the \( \hat{\Psi}_{\text{ml}} \) in a setting with linear models for \( Y \mid A, Z, C \) and \( Z \mid A, C \) (see appendix result A2.3).

### 4.2 Results

Estimation and inference were performed using the \texttt{piieffect} function implemented in the \texttt{frontdoorpiie} R package (Fulcher, 2017). Under simple linear models for the outcome and mediator variables, the variance estimator of the MLE admits a simple closed form expression (see appendix section A2.3). The variance estimator for the semiparametric estimators is described in appendix section A2.4. Alternatively, one may use the nonparametric bootstrap for inference.

In both Figure 2 and Table 1, the maximum likelihood estimator \( \hat{\Psi}_{\text{ml}} \) was only consistent under correct model specification (a) whether or not there was unmeasured confounding of the exposure-mediator relationship (b). This confirms our theoretical result as the PIIE is in fact empirically identified even if the exposure-outcome relationship is subject to unmeasured...
confounding. The MLE is not robust to model misspecification of the form in scenarios (c) and (d). On the other hand, the doubly-robust semiparametric estimator \( \hat{\Psi}_{dr} \) appears to be consistent under all scenarios (a)-(d). The semiparametric estimator \( \hat{\Psi}_1 \) which only depends on the choice of model for the density for \( Z|A, C \) has small bias in scenarios (a), (b), and (c). The semiparametric estimator \( \hat{\Psi}_2 \) which only depends on a model for the mean \( Y|A, Z, C \) and \( A|C \) has small bias in scenario (a), (b), and (d). As expected, the maximum likelihood estimator is more efficient than the semiparametric estimators when all parametric models are correctly specified. For correctly specified models, Monte Carlo coverage of 95% confidence intervals was close to nominal level. Confidence intervals based on inconsistent estimators had incorrect coverage.

Figure 2: Population intervention indirect effect by estimator and model specifications
Table 1: Operating characteristics by model specifications and estimator

| Model Specifications | \( \hat{\Psi} \) | \( \hat{\text{PIIE}} \) | Variance | Proportion bias | .95 CI | Coverage |
|----------------------|-----------------|-----------------|----------|-----------------|-------|----------|
| \( M_{y,z,a} \)      | MLE             | -18.19          | -4.61    | 0.50            | < 0.01| 0.95     |
|                      | SP 1            | -18.20          | -4.59    | 0.54            | < 0.01| 0.95     |
|                      | SP 2            | -18.19          | -4.61    | 0.60            | < 0.01| 0.95     |
|                      | SP DR           | -18.19          | -4.61    | 0.56            | < 0.01| 0.95     |
| \( M'_{y,z,a} \)     | MLE             | -18.20          | -4.59    | 0.50            | < 0.01| 0.95     |
|                      | SP 1            | -18.21          | -4.57    | 0.54            | < 0.01| 0.95     |
|                      | SP 2            | -18.20          | -4.59    | 0.55            | < 0.01| 0.94     |
|                      | SP DR           | -18.20          | -4.59    | 0.55            | < 0.01| 0.94     |
| \( M_z \)            | MLE             | -19.64          | -3.17    | 0.27            | -0.31 | 0.24     |
|                      | SP 1            | -18.23          | -4.58    | 0.54            | < 0.01| 0.95     |
|                      | SP 2            | -16.69          | -6.12    | 1.06            | 0.33  | 0.70     |
|                      | SP DR           | -18.23          | -4.58    | 0.52            | < 0.01| 0.95     |
| \( M_{y,a} \)        | MLE             | -14.15          | -8.65    | 1.61            | 0.88  | 0.12     |
|                      | SP 1            | -12.72          | -10.07   | 3.33            | 1.19  | 0.10     |
|                      | SP 2            | -18.21          | -4.59    | 0.60            | < 0.01| 0.95     |
|                      | SP DR           | -18.21          | -4.59    | 0.57            | < 0.01| 0.95     |

Note: for the \( \hat{\Psi} \) column, MLE refers to using the \( \hat{\Psi}_{mle} \) estimator for the \( \hat{\text{PIIE}} \). Likewise, SP1 refers to using \( \hat{\Psi}_1 \), SP2 refers to using \( \hat{\Psi}_2 \), and SP DR refers to using \( \hat{\Psi}_{dr} \).
5 Safer Deliveries Program in Zanzibar, Tanzania

The Safer Deliveries program aimed to reduce the high rates of maternal and neonatal mortality in Zanzibar, Tanzania by increasing the number of pregnant women who deliver in a health care facility and attend prenatal and postnatal check-ups. As of May 2017, the program was active in six (out of 11) districts in Zanzibar on the islands of Unguja and Pemba. The program trains community health workers (CHWs) selected by the Ministry of Health to participate in the program based on their literacy, expressed commitment to the improvement of health, and respectability in their communities.

The CHWs work with community leaders and staff at nearby health facilities to identify and register pregnant women and are expected to visit the woman in her home three times during pregnancy to screen for danger signs and provide counseling to help the woman prepare for a facility delivery. During the registration visit, the mobile app calculated a woman’s risk category (low, medium, or high) based on a combination of obstetric and demographic factors. Women categorized as high risk were instructed to deliver at a referral hospital. The app then calculated a recommended savings amount based on the women’s recommended delivery location. On average, high risk women were recommended to save more money than low or medium risk women as they were recommended to deliver at referral hospitals of which there are only four on the island. This analysis assessed the effectiveness of this tailored savings recommendation by risk category on actual savings.

We considered high risk category (vs. low or medium risk) as our binary exposure of interest; although, we could have also done this analysis with exposure as categorical variable. The mediator variable was recommended savings in Tanzanian Shilling (TZS), which was calculated during the first visit. The outcome variable was actual savings achieved by the woman and her family at time of her delivery. In the analysis, we adjusted for district of residence to account for regional differences in health-seeking behavior and accessibility of health facilities. The population intervention indirect effect was the best estimand for this research question as we were interested in the mediated effect of savings recommendations under the risk categories observed in the current population. Additionally, there was likely unmeasured confounding between the exposure (high risk) and outcome (actual savings) relationship because most socio-economic factors and health-seeking behavior that may be associated with other factors related to risk category and a woman’s ability to save were not collected by the program. Furthermore, confounding of exposure-mediator and mediator-outcome associations was of less concern as the app calculated the recommended savings based on the delivery location which is determined both by risk category and distance to the appropriate health facility. That is, women who are in a low risk category are recommended
to deliver at the facility closest to them, whereas women in the high risk category are recommended to deliver at one of four available referral facilities in Zanzibar.

This study included women enrolled in the Safer Deliveries program who had a live birth by May 31, 2017 (n=4,511). We excluded: 253 women from the newly-added Mkoani district of Pemba Island, 2 women with missing LMP date and EDDs, 31 women with invalid enrollment times, and 123 women with missing risk category, district, or savings information. Our final study population included 4,102 women. Therefore, the following analyses are only valid under an assumption that data are missing completely at random. The observed average savings at time of delivery was $14.09. Note that for ease of interpretation we converted from Tanzanian Shilling (1 USD = 2,236.60 TZS on May 31, 2017). We estimated the population intervention indirect effect; that is, the difference in average savings between the current population of women and a population of women had possibly contrary to the fact every woman received the savings recommendation of a low or medium risk woman. To estimate the population intervention indirect effect we employed our four estimators under the following parametric models:

\[
\text{highrisk} = \alpha_0 + \alpha_2 \text{district} + \varepsilon_a \\
\text{savings}_{\text{rec}} = \beta_0 + \beta_1 \text{highrisk} + \beta_2 \text{district} + \varepsilon_z \\
\text{savings}_{\text{act}} = \theta_0 + \theta_1 \text{highrisk} + \theta_2 \text{savings}_{\text{rec}} + \theta_3 \text{district} + \varepsilon_y
\]

The maximum likelihood estimator, \(\hat{\Psi}_{\text{mle}}\), estimated the average savings among women whose recommended savings are set to the amount they would have been recommended to save had they not been high risk to be $13.87 resulting in a PIIE of $0.22 with a 95% CI of ($0.15, $0.30). The semiparametric estimator that only includes models for \(A|C\) and \(Y|A, Z, C\), \(\hat{\Psi}_2\), gave almost identical results. On the other hand, the doubly robust semiparametric estimator estimated the PIIE to be $13.95 with 95% CI of (-$0.03, $0.32). The semiparametric estimator that only depends on a parametric model for \(Z|A, C\), \(\hat{\Psi}_1\) resulted in very similar inferences to the doubly-robust estimator. We may infer that the parametric model for \(Y | Z, A, C\) was likely misspecified as the maximum likelihood estimator does not posit a model for \(A|C\) and \(\hat{\Psi}_{\text{mle}}\) and \(\hat{\Psi}_2\) align almost exactly. Thus, there were two plausible estimators for the PIIE: SP 1 or SP DR. For both these estimators, the population intervention indirect effect was not significantly different from zero revealing that the tailored savings recommendations to high risk women may not effect their actual savings by the time of their delivery. On average, if high risk women had been recommended to save what they would have if they were low to medium risk, there would not have been a significant change in their savings behavior.
Table 2: Effect of risk category on actual savings mediated by recommended savings ($n = 4,102$)

|      | $\hat{\Psi}$ | $\hat{PIIE}$ | Standard Error | 95% CI     |
|------|---------------|--------------|---------------|------------|
| MLE  | 13.87         | 0.22         | 0.04          | (0.15, 0.30) |
| SP 1 | 14.08         | 0.02         | 0.11          | (-0.20, 0.23) |
| SP 2 | 13.87         | 0.22         | 0.05          | (0.13, 0.31)  |
| SP DR| 13.95         | 0.14         | 0.09          | (-0.03, 0.32) |

6 Discussion

In this paper, we have argued that the population intervention indirect effect may be used in place of the natural indirect effect for two reasons. First, in the case of a harmful exposure, interpretation of the NIE requires conceiving of an intervention that would force persons to acquire that harmful exposure, which is unlikely to be of interest; however, it may be possible to conceive of an intervention (at least hypothetically) that might prevent a person from being exposed. Second, it is robust to unmeasured confounding of the exposure-outcome relationship. Despite the latter desirable property, the use of the PIIE does not obviate concern about unmeasured confounding of the exposure-mediator or mediator-outcome relation. When such confounding is of concern, a sensitivity analysis should be performed (Robins et al., 2000; Lin et al., 1998; VanderWeele and Arah, 2011). Investigators should exercise caution if they also wish to report the PIDE and PIE as these effects are not robust to exposure-outcome confounding. If exposure-outcome unmeasured confounding can be ruled out with reasonable certainty, then one can estimate the PIDE using our doubly-robust estimator for $\Psi$ and the well-known doubly-robust estimator for $E(Y(a^*))$ from Robins et al. (2000). Likewise, the PIE can be estimated using the doubly robust estimator developed by Hubbard and Van der Laan (2008).

Despite the front-door criterion being available in the literature for several years, this is the first methodology developed for semiparametric estimation and inference of the front-door functional $\Psi$. Therefore, when an investigator believes she has identified one or more mediator variables that satisfy the front-door criterion, she can use our proposed methodology to obtain an estimate of the PIE or the average causal effect that is not only doubly-robust, but also robust to unmeasured confounding of the exposure-outcome relation.
References

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. *Department of Statistics, UCLA*, 2005.

P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 2. Springer New York, 1998.

G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2017.

A. M. Cotter, A. G. Garcia, M. L. Duthely, B. Luke, and M. J. O’Sullivan. Is antiretroviral therapy during pregnancy associated with an increased risk of preterm delivery, low birth weight, or stillbirth? *The Journal of infectious diseases*, 193(9):1195–1201, 2006.

A. B. Ettinger and C. E. Argoff. Use of antiepileptic drugs for nonepileptic conditions: psychiatric disorders and chronic pain. *Neurotherapeutics*, 4(1):75–83, 2007.

I. Fulcher. *frontdoorpiie*. https://github.com/isabelfulcher/frontdoorpiie, 2017.

L. B. Holmes, E. A. Harvey, B. A. Coull, K. B. Huntington, S. Khoshbin, A. M. Hayes, and L. M. Ryan. The teratogenicity of anticonvulsant drugs. *New England Journal of Medicine*, 344(15):1132–1138, 2001.

A. E. Hubbard and M. J. Van der Laan. Population intervention models in causal inference. *Biometrika*, 95(1):35–47, 2008.

K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

D. Y. Lin, B. M. Psaty, and R. A. Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, pages 948–963, 1998.

E. S. Machado, C. B. Hofer, T. T. Costa, S. A. Nogueira, R. H. Oliveira, T. F. Abreu, L. A. Evangelista, I. F. Farias, R. T. Mercadante, L. G. Maria de Fatima, et al. Pregnancy outcome in hiv-1 infected women receiving combination antiretroviral therapy prior versus after conception. *Sexually transmitted infections*, 2008.
F. Martin and G. P. Taylor. Increased rates of preterm delivery are associated with the initiation of highly active antiretroviral therapy during pregnancy: a single-center cohort study. *The Journal of Infectious Diseases*, 196(4):558–561, 2007.

W. K. Newey and J. M. Robins. Cross-fitting and fast remainder rates for semiparametric estimation. 2017.

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

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

J. Pearl. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–436, 2012.

J. Robins. 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, 1986.

J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

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

E. J. Tchetgen Tchetgen and K. Phiri. Evaluation of medication-mediated effects in pharmacoepidemiology. *Epidemiology (Cambridge, Mass.)*, 28(3):439, 2017.

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

T. J. VanderWeele and O. A. Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, 22(1):42–52, 2011.

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

A1 Proofs of lemmas and theorems

Proof of Lemma 1. Generalized front-door functional derivation

\[ \Psi = E[Y(Z(a^*))] \]
\[ = \sum_{c,a,z} E(Y(a, Z(a^*))|Z(a^*) = z, A = a, C = c)Pr(Z(a^*) = z|A = a, C = c)Pr(A = a, C = c) \]
\[ M_1 = \sum_{c,a,z} E(Y(a, z)|Z(a^*) = z, A = a, C = c)Pr(Z(a^*) = z|A = a, C = c)Pr(A = a, C = c) \]
\[ M_2 = \sum_{c,a,z} E(Y(a, z)|Z(a^*) = z, A = a, C = c)Pr(Z(a^*) = z|A = a^*, C = c)Pr(A = a, C = c) \]
\[ M_1, M_3 = \sum_{c,a,z} E(Y(a, z)|a = a, C = c)Pr(Z = z|A = a^*, C = c)Pr(A = a, C = c) \]
\[ M_4 = \sum_{c,a,z} E(Y(a, z)|Z = z, A = a, C = c)Pr(Z = z|A = a^*, C = c)Pr(A = a, C = c) \]
\[ M_1 = \sum_{c,a,z} E(Y(Z = z, A = a, C = c)Pr(Z = z|A = a^*, C = c)Pr(A = a, C = c) \]
\[ = \sum_{z,c} Pr(Z = z|A = a^*, C = c) \sum_a E(Y|Z = z, A = a, C = c)Pr(A = a|C = c)Pr(C = c) \]

Proof of Lemma 2.

\[ E[Y \frac{f(Z|a^*, C)}{f(Z|A, C)}] = \sum_{y,a,z,c} y \frac{f(z|a,c)}{f(z|a,c)} f(y|a, z, c) f(z|a^*, C) f(a|c) f(c) \]
\[ = \sum_{a,z,c} E(Y|a, z, c) f(z|a^*, c) f(a|c) f(c) \]
\[ = \Psi \]

The proof of asymptotic normality is fairly standard under the usual regularity conditions once unbiasedness of the estimating equation is established (see Theorem 1A in Robins, Mark, and Newey (1992)).
Proof of Lemma 3.

\[
E \left[ \frac{I(A = a^*)}{f(A|C)} E \left( E \left\{ Y | A, Z, C \right\} | C \right) \right] = \sum_{z, a, c} \frac{I(\bar{a} = a^*)}{f(z|a, c)} f(c) E \left( E \left\{ Y | a, z, c \right\} | c \right)
\]

\[
= \sum_{z, c} f(z|a^*, c) f(c) \sum_a E \left( Y | a, z, c \right) f(a|c)
\]

\[
= \Psi
\]

The proof of asymptotic normality is fairly standard under the usual regularity conditions once unbiasedness of the estimating equation is established (see Theorem 1A in Robins, Mark, and Newey (1992)).

Proof of Theorem 1. Efficient influence function derivation

We aim to find an efficient influence function, \( \varphi^{eff}(Y, Z, A, C) \), for \( \Psi = E[Y(Z(a^*))] \) under model corresponding to Figure 1c. Our functional is nonparametrically identified under the causal model represented by a complete graph. In other words, the causal model induces no restrictions on the observed data. Thus, there is a unique influence function, \( \varphi^{eff}(Y, Z, A, C) \), and it achieves the semiparametric efficient bound of \( \Psi \) in \( \mathcal{M}_{np} \). We will use the definition of pathwise differentiability to find the efficient influence function.

\[
\frac{d}{dt} \Psi(F_t) = E[\varphi^{eff}(Y, Z, A, C) \times S(Y, A, Z, C)]
\]

where \( S(Y, A, Z, C) \) is the score corresponding to the whole model.
\[
\frac{d}{dt} \Psi(F_t) = \sum_{z,a,c} \frac{d}{dt} E_t[Y|A = a, Z = z, C = C] f_t(z|a^*, c) f_t(a|c) f_t(c)
\]

(from now on, for convenience I will just use \(f(.)\))

\[
= \sum_{z,a,c} \sum_y y \frac{d}{dt} (f(y|a, z, c) f(z|a^*, c) f(a|c) f(c))
\]

\[
= \sum_{z,a,c} \sum_y y S(y|a, z, c) f(y|a, z, c) f(z|a^*, c) f(a|c) f(c)
\]

\[
+ \sum_{z,a,c} E[Y|a, z, c] S(z|a^*, c) f(z|a^*, c) f(a|c) f(c)
\]

\[
+ \sum_{z,a,c} E[Y|a, z, c] f(z|a^*, c) S(a|c) f(a|c) f(c)
\]

\[
+ \sum_{z,a,c} E[Y|a, z, c] f(z|a^*, c) f(a|c) S(c) f(c)
\]

...detail for each of the four terms portion given below...

\[
= E[(Y - E(Y|A, Z, C)) \frac{f(Z|a^*, C)}{f(Z|A, C)} \times S(Y, A, Z, C)] \tag{A1}
\]

\[
+ E\left[ \left( \sum_a E[Y|a, Z, C] f(a|C) \right. \right.
\]

\[
\left. - \sum_{a, \bar{m}} E(Y|a, \bar{m}, C) f(\bar{m}|A, C) f(a|C) \right) \frac{I(\bar{A} = a^*)}{f(A|c)} \times S(Y, A, Z, C) \right] \tag{A2}
\]

\[
+ E\left[ \left( \sum_m E[Y|A, z, C] f(z|a^*, C) \right.ight.
\]

\[
\left. - \sum_{a, m} E[Y|a, z, C] f(z|a^*, C) f(a|C) \right) \times S(Y, A, Z, C) \right] \tag{A3}
\]

\[
+ E\left[ \left( \sum_{z,a} E[Y|a, z, C] f(z|a^*, C) f(a|C) - \Psi \right) \times S(Y, A, Z, C) \right] \tag{A4}
\]

Each of the four terms (A1)-(A4) will be handled in turn. The goal is to get them in the form

\[
E[IF \times S] = \sum_{i=1}^{4} E[IF_i \times S(Y, A, Z, C)].
\]
(A1) = \sum_{z,a,c} \sum_{y} yS(y|a, z, c)f(y|a, z, c)f(z|a^*, c)f(a|c)f(c)

= \sum_{z,a,c,y} y \frac{f(z|a^*, c)}{f(z|a, c)} f(z|a, c)f(y|a, z, c)f(a|c)f(c)S(y|a, z, c)

\approx \sum_{z,a,c,y} (y - E[Y|a, z, c]) \frac{f(z|a^*, c)}{f(z|a, c)} f(y|a, z, c)f(z|a, c)f(a|c)f(c)

\times (S(y|a, z, c) + S(z|a, c) + S(a|c) + S(c))

= \sum_{z,a,c,y} (y - E[Y|a, z, c]) \frac{f(z|a^*, c)}{f(z|a, c)} f(y|a, z, c)f(y, a, z, c)

= E[(Y - E[Y|A, Z, C]) \frac{f(Z|a^*, C)}{f(Z|A, C)} \times S(Y, A, Z, C)]

* The equality will hold because the added term will evaluate to zero as the expectation of a score is zero (in brackets),

\sum_{z,a,c} E[Y|a, z, c] \frac{f(z|a^*, c)}{f(z|a, c)} f(z|a, c)f(a|c)f(c) \left[ \sum_{y} S(y|a, z, c)f(y|a, z, c) \right] = 0

** Similar to above, the additional terms will all evaluate to zero as the term in the large brackets is zero:

\sum_{z,a,c} \frac{f(z|a^*, c)}{f(z|a, c)} f(z|a, c)f(a|c)f(c)S(z|a, c) \left[ \sum_{y} (y - E[Y|a, z, c])f(y|a, z, c) \right] = 0

\sum_{z,a,c} \frac{f(z|a^*, c)}{f(z|a, c)} f(z|a, c)f(a|c)f(c)S(a|c) \left[ \sum_{y} (y - E[Y|a, z, c])f(y|a, z, c) \right] = 0

\sum_{z,a,c} \frac{f(z|a^*, c)}{f(z|a, c)} f(z|a, c)f(a|c)f(c)S(c) \left[ \sum_{y} (y - E[Y|a, z, c])f(y|a, z, c) \right] = 0
\[
(A2) = \sum_{z,a,c} E[Y|a, z, c] S(z|a^*, c) f(z|a^*, c) f(a|c) f(c)
\]

\[
= \sum_{z,c} \sum_{\bar{a}} \left( \sum_a E[Y|a, z, c] f(a|c) \right) I(\bar{a} = a^*) S(z|\bar{a}, c) f(z|\bar{a}, c) f(c)
\]

\[
= \sum_{z,c,\bar{a}} \left( \sum_a E[Y|a, z, c] f(a|c) \right) \frac{I(\bar{a} = a^*)}{f(\bar{a}|c)} S(z|\bar{a}, c) f(z|\bar{a}, c) f(\bar{a}|c) f(c)
\]

\[
= \sum_{z,c,\bar{a}} \left( \sum_a E[Y|a, z, c] f(a|c) - \sum_{a,\bar{m}} E(Y|a, \bar{m}, c) f(\bar{m}|\bar{a}, c) f(a|c) \right)
\]

\[
\times \frac{I(\bar{a} = a^*)}{f(\bar{a}|c)} f(z|\bar{a}, c) f(\bar{a}|c) f(c)
\]

\[
= \sum_{z,c,\bar{a},y} \left( \sum_a E[Y|a, z, c] f(a|c) - \sum_{a,\bar{m}} E(Y|a, \bar{m}, c) f(\bar{m}|\bar{a}, c) f(a|c) \right)
\]

\[
\times \frac{I(\bar{a} = a^*)}{f(\bar{a}|c)} f(y, \bar{a}, z, c) S(y, a, z, c)
\]

\[
= E\left[ \sum_a E[Y|a, Z, C] f(a|C) - \sum_{a,\bar{m}} E(Y|a, \bar{m}, C) f(\bar{m}|\bar{A}, C) f(a|C) \right] \frac{I(\bar{A} = a^*)}{f(\bar{A}|c)} \times S(Y, A, Z, C)
\]

*The reasoning here is identical to that for the first term.

**The reasoning here is identical to that for the first term.
Thus, the efficient influence function under the nonparametric model is as follows:

\[
\varphi^{eff}(Y, Z, A, C) = E \left[ (Y - E(Y|A, Z, C)) \frac{f(Z|a^*, C)}{f(Z|A, C)} \right] \\
+ I(A = a^*) \frac{E[Y|a, Z, C]f(a|C) - E(Y|a, \bar{z}, C) f(\bar{z}|A, C) f(a|C)}{f(A|C)} \\
+ \sum_z E[Y|A, z, C] f(z|a^*, C) - \Psi
\]
Proof of Theorem 2.

We first show that the influence function derived in Theorem 1 has expectation 0 if one of the following scenarios holds:

1. $E(Y|a, z, c)$ and $f(a|c)$ are correct
2. $f(z|a, c)$ is correct

1. $E(Y|a, z, c)$ & $f(a|c)$ correctly specified and $	ilde{f}(z|a, c)$ misspecified

$$E[\phi^e_{\tilde{f}}] = E[(Y - E(Y|A, Z, C)) \frac{\tilde{f}(Z|a^*, C)}{f(Z|A, C)}]$$

$$+ E[I(A = a^*) \frac{1}{f(A|C)} \left( \sum_a E[Y|a, Z, C] f(a|C) - \sum_{a, z} E(Y|a, z, C) \tilde{f}(z|A, C) f(a|C) \right)]$$

$$+ E[\sum_z E[Y|A, z, C] \tilde{f}(z|a^*, C) - \Psi]$$

$$= 0 + \sum_{a', z, c} I(a' = a^*) \frac{1}{f(a'|c)} \left( \sum_a E[Y|a, z, c] f(a|c) - \sum_{a, z} E(Y|a, z, C) \tilde{f}(z|a', c) f(a|c) \right) f(z, a', c)$$

$$+ E[\sum_z E[Y|A, z, C] \tilde{f}(z|a^*, C) - \Psi]$$

$$= \sum_{z,c} \sum_a E[Y|a, z, c] f(a|c) f(z|a^*, c) f(c) - \sum_{a, z, c} E(Y|a, z, C) \tilde{f}(z|a^*, c) f(a|c) f(c)$$

$$+ E[\sum_z E[Y|A, z, C] \tilde{f}(z|a^*, C) - \Psi]$$

$$= \Psi - \sum_{a, z, c} E(Y|a, z, C) \tilde{f}(z|a^*, c) f(a|c) f(c) + \sum_{a, c} \sum_z E[Y|a, z, c] \tilde{f}(z|a^*, c) f(a|c) f(c) - \Psi$$

$$= 0$$
2. $f(z|a,c)$ correctly specified and $\bar{E}(Y|a,z,c)$ & $\bar{f}(a|c)$ misspecified

$$E[\phi_{\text{eff}}] = E[(Y - \bar{E}(Y|A,Z,C))(\frac{f(Z|a^*,C)}{f(Z|A,C)})$$

$$+ E[\frac{I(A = a^*)}{f(A|C)}(\sum_a \bar{E}[Y|a,Z,C]\bar{f}(a|C) - \sum_a,\bar{z} \bar{E}(Y|a,z,C) f(z|A,C)\bar{f}(a|C))]$$

$$+ E[\sum_z \bar{E}[Y|A,z,C] f(z|a^*,C) - \Psi]$$

$$= \sum_{c,a,z} (E(Y|a,z,c) - \bar{E}(Y|a,z,c)) f(z|a^*,c) f(a|c) f(c)$$

$$+ \sum_{c,a,z} \frac{1}{\bar{f}(a|c)}(\bar{E}[Y|a,z,c] \bar{f}(a|c) f(z|a^*,c) f(a^*|c) f(c)$$

$$- \bar{E}[Y|a,z,c] \bar{f}(a|c) f(z|a^*,c) f(a^*) f(c)$$

$$+ \sum_{c,a,z} \bar{E}[Y|a,z,c] f(z|a^*,c) f(a|c) f(c) - \Psi$$

$$= \Psi - \sum_{c,a,z} \bar{E}[Y|a,z,c] f(z|a^*,c) f(a|c) f(c) + \sum_{c,a,z} \bar{E}[Y|a,z,c] f(z|a^*,c) f(a|c) f(c) - \Psi$$

$$= 0$$

Assuming the regularity conditions of Theorem 1A in Robins, Mark, and Newey (1992) hold for $\varphi_{\text{eff}}(Y,Z,A,C)$, the expression follows by standard Taylor expansion arguments:

$$\sqrt{n}(\hat{\Psi}_{dr} - \Psi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\text{eff}}(Y_i,Z_i,A_i,C_i) + o_p(1)$$

The asymptotic distribution of the left hand side under $\mathcal{M}_{union}$ follows from the previous equation by the Central Limit Theorem and Slutsky’s.
A2 Additional materials

A2.1 Judea Pearl’s front-door criterion

\[ E(Y(a^*)) = \sum_{z,c} E(Y(a^*)|Z(a^*) = z, C = c)Pr(Z(a^*) = z|C = c)Pr(C = c) \]

\[ \overset{M2}{=} \sum_{z,c} E(Y(a^*)|Z(a^*) = z, C = c)Pr(Z(a^*) = z|C = c, A = a^*)Pr(C = c) \]

\[ \overset{M1}{=} \sum_{z,c} E(Y(a^*, z)|Z(a^*) = z, C = c)Pr(Z = z|C = c, A = a^*)Pr(C = c) \]

\[ \overset{F1}{=} \sum_{z,c} E(Y(z)|Z(a^*) = z, C = c)Pr(Z = z|C = c, A = a^*)Pr(C = c) \]

\[ \overset{M3}{=} \sum_{z,c} E(Y(z)|C = c)Pr(Z = z|C = c, A = a^*)Pr(C = c) \]

\[ = \sum_{z,c} \left[ \sum_{a} E(Y(z)|C = c, A = a)Pr(A = a|C = c) \right] Pr(Z = z|C = c, A = a^*)Pr(C = c) \]

\[ = \sum_{z,c} Pr(Z = z|C = c, A = a^*) \sum_{a} E(Y(z)|A = a, C = c)Pr(A = a|C = c)Pr(C = c) \]

\[ \overset{M4}{=} \sum_{z,c} Pr(Z = z|C = c, A = a^*) \sum_{a} E(Y(z)|A = a, Z = z, C = c)Pr(A = a|C = c)Pr(C = c) \]

\[ \overset{M1}{=} \sum_{z,c} Pr(Z = z|C = c, A = a^*) \sum_{a} E(Y|A = a, Z = z, C = c)Pr(A = a|C = c)Pr(C = c) \]

\[ = \Psi \]

A2.2 NPSEM-IE Interpretation of the causal diagram

We can formalize the conditions for identification of \( \Psi \) under Figure 1c or assumptions M1-M4 using a system of equations known as “Nonparametric Structural Equation Model”. We assign a system of equations for each variable as below:

\[ U = g_U(\varepsilon_U) \]
\[ C = g_C(\varepsilon_C) \]
\[ A = g_A(C, U, \varepsilon_A) \]
\[ Z = g_Z(A, C, \varepsilon_Z) \]
\[ Y = g_Y(Z, A, U, C, \varepsilon_Y) \]

Each of the five random variables on this graph are associated with a distinct, arbitrary function, denoted \( g \), and a distinct random disturbance, denoted \( \varepsilon \), each with a subscript corresponding to
its respective random variable. Each variable is generated by its corresponding function, which
depends only on all variables that affect it directly. These equations provide a nonparametric
algebraic interpretation of the Figure (1c), and are helpful in defining potential outcomes. The
identification conditions given above can be formalized in terms of independence conditions about
the errors; specifically, we require all the errors to be independent.

A2.3 Parametric derivation for PIIE

Here, we build a parametric expression $E[Y(Z(a^*))]$ where we include parametric models for both
$Y$ and $Z$. Due to the fact $Y$ and $Z$ are continuous, we do not need to specify a model for $A$. We
will compare the parametric estimator $\hat{\Psi}_{mle}$ to the semiparametric estimators $\hat{\Psi}_1, \hat{\Psi}_2, \hat{\Psi}_{dp}$ in our
simulation study. The two models are as follows:

$$
E[Y|A = a, Z = z, C = c] = \theta_0 + \theta_1 a + \theta_2 z + \theta_3 az + \theta_4^T c
$$

$$
E[Z|A = a^*, C = c] = \beta_0 + \beta_1 a^* + \beta_2^T c
$$
\[
E[Y(Z(a^*))] = \sum_c \sum_z Pr(Z = z | A = a^*, C = c) \sum_a E[Y | A = a, Z = z, C = c] Pr(A = a, C = c)
\]
\[
= \sum_c \sum_z Pr(Z = z | A = a^*, C = c) \sum_a (\theta_0 + \theta_1 a + \theta_2 z + \theta_3 a z + \theta_4^T c) Pr(A = a, C = c)
\]
\[
= \sum_c Pr(C = c) \sum_z Pr(Z = z | A = a^*, C = c) \times
\]
\[
(\theta_0 + \theta_1 E[A|C = c] + \theta_2 z + \theta_3 a z + \theta_4^T c)
\]
\[
= \theta_0 + \theta_1 E[A|C = c] Pr(C = c) + \theta_2 \sum_c Pr(C = c) \sum_z z Pr(Z = z | A = a^*, C = c)
\]
\[
+ \theta_3 \sum_c Pr(C = c) \sum_z z E[A|C = c] Pr(Z = z | A = a^*, C = c) + \theta_4^T E[C]
\]
\[
= \theta_0 + \theta_1 E[A] + \theta_2 \sum_c Pr(C = c) E[A|C = c] E[Z | A = a^*, C = c] + \theta_4^T E[C]
\]
\[
+ \theta_3 \sum_c Pr(C = c) E[A|C = c] (\beta_0 + \beta_1 a^* + \beta_2^T c)
\]
\[
+ \theta_3 \sum_c Pr(C = c) E[A|C = c] (\beta_0 + \beta_1 a^* + \beta_2^T c) + \theta_4^T E[C]
\]
\[
= \theta_0 + \theta_1 E[A] + \theta_2 \beta_0 + \theta_2 \beta_1 a^* + \theta_2 \beta_2^T E[C] + \theta_3 \beta_0 E[A] + \theta_3 \beta_1 a^* E[A]
\]
\[
+ \theta_3 \beta_2^T \sum_c c Pr(C = c) E[A|C = c] + \theta_4^T E[C]
\]
\[
= \theta_0 + \theta_1 E[A] + \theta_2 \beta_0 + \theta_2 \beta_1 a^* + \theta_2 \beta_2^T E[C] + \theta_3 \beta_0 E[A] + \theta_3 \beta_1 a^* E[A]
\]
\[
+ \theta_3 \beta_2^T E[CE[A|C = c]] + \theta_4^T E[C]
\]
\[
= \theta_0 + \theta_2 \beta_0 + \theta_2 \beta_1 a^* + (\theta_1 + \theta_3 \beta_0 + \theta_3 \beta_1 a^*) E[A] + (\theta_2 \beta_2^T + \theta_4^T) E[C] + \theta_3 \beta_2^T E[AC]
\]

In this setting, there is a closed form expression for the \(\theta\). In settings where the outcome or mediator variable are binary, the variance can be computed using the sandwich variance or via the nonparametric bootstrap.

\[
Var(PIIE) = \beta_1 \beta_2 (\beta_1 \beta_3 \text{Cov}(A, A^2) + \beta_1 \beta_2 \text{Var}(A)) + \beta_1 \beta_3 (\text{Var}(A^2) \beta_1 \beta_3 + \beta_1 \beta_2 \text{Cov}(A, A^2))
\]
\[
+ (E(A) \beta_2 + E(A^2) \beta_3) \text{Var}(\beta_1) + E(A) \beta_1 (E(A) \beta_1 \text{Var}(\beta_2) + E(A^2) \beta_1 \text{Var}(\beta_2, \beta_3))
\]
\[
+ E(A^2) \beta_1 (E(A) \beta_1 \text{Cov}(\beta_2, \beta_3) + E(A^2) \beta_1 \text{Var}(\beta_3))
\]

For estimation with binary \(A\), \(E(A) = E(A^2) = \overline{A}\), \(Var(A) = Var(A^2) = \text{Cov}(A, A^2) = S_A^2\) (sample variance), and all the parameters are estimated via their MLE in R.
A2.4 Sandwich variance

Let $\theta$ denote the vector of all $K$ parameters and $U(\theta) = [U_T^1, ..., U_T^K]^T$ denote the score vector where the $K$th score corresponds to the score for $\Psi$. A consistent estimator for the asymptotic variance of $\theta$:

$$\hat{\text{Var}}(\theta) = \left( \sum_{i=1}^{n} \frac{dU(\theta)}{d\theta}|_{\theta=\hat{\theta}} \right)^{-1} U(\hat{\theta})^T \left( \sum_{i=1}^{n} \frac{dU(\theta)}{d\theta}|_{\theta=\hat{\theta}} \right)^{-1}$$

Further, a consistent estimator for the asymptotic variance of $\hat{\Psi}$ will correspond to the $\hat{\text{Var}}(\theta)_{k,k}$ element.

A2.5 Population intervention effect and the total effect among exposed

For binary $A$ and $a^* = 0$,

$$PIE(0) = E(Y - Y(0))$$
$$= E(AY + (1 - A)Y - Y(0))$$
$$= E(A(AY(1) + (1 - A)Y(0) - Y(0))) \text{ by Consistency}$$
$$= E(AY(1) - Y(0)) + Y(0) - Y(0))$$
$$= E(E(AY(1) - Y(0)|A = a))$$
$$= E(Y(1) - Y(0)|A = 1))Pr(A = 1)$$
$$= ETT \ Pr(A = 1)$$

A2.6 Alternative population intervention effect decomposition

We could have used an alternative decomposition of the population intervention effect,

$$PIE(a^*) = E[Y(A, Z(A)) - Y(a^*, Z(a^*))] = E[I_{a^*}(a^*, Z)] + E[I_{a^*}(a^*, Z)]$$

The use of this alternative decomposition would not guarantee robustness to exposure-outcome confounding as the indirect effect includes the term $E(Y(a^*))$, which requires no unmeasured confounding of exposure-outcome relation for identification (similar to our PIDE). Additionally, identification of the term $E(Y(a^*, Z))$ requires a different set of assumptions that will not lead to a connection with the frontdoor formula. If we were to condition on the exposed, the indirect and direct effects from (A5) aligns with the effect decomposition of the total effect on the exposed (ETT) described by Vansteelandt and Vanderweele (2012),
$$ETT = E(Y(1) - Y(0)|A = 1) = E[Y(0, Z) - Y(0)|A = 1] + E[Y - Y(0, Z)|A = 1]$$

See appendix section A2.6 for the expression of the population intervention effect as a function of the ETT. The identification conditions needed for the Vansteelandt and Vanderweele (2012)'s indirect effect are different than those needed for the PIIE. These are listed in their paper (section 4), but we also state them using our notation below:

M1. Consistency assumptions: (1) If \( A = a \), then \( Z(a) = Z \) w.p.1,
(2) If \( A = a \), then \( Y(a) = Y \) w.p.1,
(3) If \( A = a \) and \( Z = z \), then \( Y(a, z) = Y \) w.p.1

M4. \( Y(a, z) \perp Z|A = a, C = c \) \( \forall z, a, c \)

M5. \( Y(a, z) \perp A|C = c \) \( \forall z, a, c \)

A2.7 Simulation results (color version)