Harvard University
Harvard University Biostatistics Working Paper Series

Year 2015 *Paper* 196

On Partial Identification of the Pure Direct Effect

Caleb Miles[*] Phyllis Kanki[†]
Seema Meloni[‡] Eric Tchetgen Tchetgen[**]

[*]University of California - Berkeley, calebmiles@gmail.com
[†]Harvard T.H. Chan School of Public Health, pkanki@hsph.harvard.edu
[‡]Harvard T.H. Chan School of Public Health
[**]Harvard T.H. Chan School of Public Health, etchetge@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper196

On Partial Identification of the Pure Direct Effect

Caleb Miles, Phyllis Kanki, Seema Meloni, and Eric Tchetgen Tchetgen*

Abstract

In causal mediation analysis, nonparametric identification of the pure (natural) direct effect typically relies on, in addition to no unobserved pre-exposure confounding, fundamental assumptions of (i) so-called "cross-world-counterfactuals" independence and (ii) no exposure-induced confounding. When the mediator is binary, bounds for partial identification have been given when neither assumption is made, or alternatively when assuming only (ii). We extend existing bounds to the case of a polytomous mediator, and provide bounds for the case assuming only (i). We apply these bounds to data from the Harvard PEPFAR program in Nigeria, where we evaluate the extent to which the effects of antiretroviral therapy on virological failure are mediated by a patient's adherence, and show that inference on this effect is somewhat sensitive to model assumptions.

Keywords: Cross-world counterfactual, Mediation, Natural direct effect, Partial identification, Pure direct effect, Single World Intervention Graph

*Caleb Miles is Postdoctoral Fellow, Department of Biostatistics, University of California, Berkeley 94720-7358. Phyllis Kanki is Professor and Seema Meloni is Research Associate, Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115. Eric Tchetgen Tchetgen is Professor, Departments of Biostatistics and Epidemiology, Harvard School of Public Health, Boston, MA 02115. The authors gratefully acknowledge the hard work and dedication of the clinical, data, and laboratory staff at the PEPFAR supported Harvard/AIDS Prevention Initiative in Nigeria (APIN) hospitals that provided secondary data for this analysis. This work was funded, in part, by grants from the U.S. National Institutes of Health. The contents are solely the responsibility of the authors and do not represent the official views of the funding institutions.

1. INTRODUCTION

Causal mediation analysis seeks to determine the role that an intermediate variable plays in transmitting the effect from an exposure to an outcome. An indirect effect refers to the effect that goes through the intermediate variable in mediation analysis; a direct effect is a measure of the effect that does not. The study of causal mediation has in recent years enjoyed an explosion in popularity (Robins and Greenland, 1992; Robins, 1999, 2003; Pearl, 2001; Avin et al., 2005; Taylor et al., 2005; Petersen et al., 2006; Ten Have et al., 2007; Albert, 2008; Goetgeluk et al., 2008; van der Laan and Petersen, 2008; VanderWeele, 2009; VanderWeele and Vansteelandt, 2009, 2010; Imai et al., 2010a,b; Albert and Nelson, 2011; Tchetgen Tchetgen, 2011; VanderWeele, 2011; Albert, 2012; Tchetgen Tchetgen and Shpitser, 2012; Wang and Albert, 2012; Shpitser, 2013; Tchetgen Tchetgen, 2013; Tchetgen Tchetgen and Shpitser, 2014; Wang et al., 2013; Albert and Wang, 2015; Hsu et al., 2015), not only in terms of theoretical developments, but also in practice, most notably in the fields of epidemiology and social sciences. This strand of work is based on ideas originating from Robins and Greenland (1992) and Pearl (2001) grounded in the language of potential outcomes (Splawa-Neyman et al., 1990; Rubin, 1974, 1978) to give nonparametric definitions of effects involved in mediation analysis, allowing for settings where interactions and nonlinearities may be present.

Consider an intervention which sets the exposure of interest for all persons in the population to one of two possible values, a reference value or an active value. The total effect of such an intervention corresponds to the change of the counterfactual outcome mean if the exposure were set to the active value compared with if it were set to the reference value. Robins and Greenland (1992) formalized the concept of effect decomposition of the total effect into direct and indirect effects by defining pure direct and indirect effects. Pearl (2001) relabeled these effects as natural direct and indirect effects. The pure direct effect (PDE) corresponds to the change in the counterfactual outcome mean under an intervention which changes a person's exposure status from the reference value to the active value, while maintaining the person's mediator to the value it would have had under the exposure reference value. In contrast, the natural indirect effect (NIE) corresponds to the change in the average counterfactual outcome under an intervention that sets a person's exposure value to the active value, while changing the value of the mediator from the

value it would have had under the reference exposure value, to its value under the active exposure value. The PDE and NIE sum to give the total effect.

Identification of these natural effects has been somewhat controversial as it requires assumptions that may be overly restrictive for many applications in the health sciences. First, identification invokes a so-called cross-world-counterfactuals-independence assumption, which by virtue of involving counterfactuals under conflicting interventions on the exposure, can neither be enforced experimentally nor tested empirically (Pearl, 2001; Robins and Richardson, 2010). Secondly, a necessary assumption for identification rules out the presence of exposure-induced confounding of the mediator's effect on the outcome, even if all confounders are observed. While this assumption is in principle testable provided no unmeasured confounding, more often than not, post-exposure covariates are altogether ignored in routine application, in which case mediation analyses may be invalid. These issues have recently been considered, and some work has been done on partial or point identification under a weaker assumption. Specifically, on the one hand Robins and Richardson (2010) and Tchetgen Tchetgen and VanderWeele (2014) provide conditions for point identification of the pure direct effect when a confounder is directly affected by the exposure. On the other hand, Robins and Richardson (2010) give bounds for the pure direct effect for binary mediator without making the cross-world-counterfactual-independence assumption, but assuming no exposure-induced confounding of the mediator-outcome relation, and Tchetgen Tchetgen and Phiri (2014) extend these bounds to account for exposure-induced confounding. Bounds are commonly employed in causal inference when structural assumptions are not sufficiently strong to give point identification of a causal parameter of interest (Robins, 1989; Balke and Pearl, 1997; Zhang and Rubin, 2003; Kaufman et al., 2005; Cheng and Small, 2006; Cai et al., 2008; Sjölander, 2009; Taguri and Chiba, 2015). We build on this previous work to provide a number of new nonparametric bounds for the pure direct effect allowing for a polytomous mediator when either (i) exposure-induced confounding is present, or (ii) one does not assume that cross-world counterfactuals of the mediating and outcome variables are independent, or (iii) both (i) and (ii) hold.

We apply these bounds to data from the Harvard PEPFAR program in Nigeria, where we evaluate the extent to which the effects of antiretroviral therapy on virological failure are mediated by a patient's adherence. We show that PEPFAR results are sensitive to the choice of

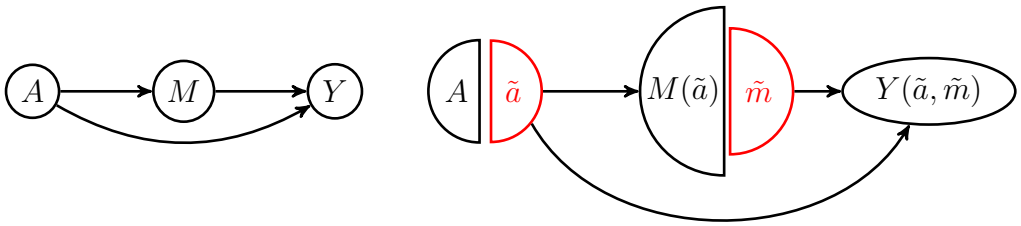

Figure 1: (a) The three-node mediation directed acyclic graph in a setting with no confounding. The nodes represent random variables, and the arrows represent possible causal effects of one random variable on another. (b) The single-world intervention graph in the setting of (a) under the intervention setting A to $\tilde{a}$ and M to $\tilde{m}$. The black nodes represent random variables under this intervention, the red nodes represent the level an intervened random variable takes under this intervention, and the arrows represent possible causal effects of one variable under this intervention on another.

assumptions made, consequently, we counsel investigators employing these effects to exercise caution in considering the basis for point identification and to explicitly state the assumptions required for them to be valid. Where assumptions are empirically untestable, they should be argued for on the basis of scientific understanding, and ideally the alternative should be explored by employing partial identification bounds given both here and elsewhere. While some work has been done to develop sensitivity analyses for unmeasured confounding of the mediator (Tchetgen Tchetgen, 2011; Tchetgen Tchetgen and Shpitser, 2012; Vansteelandt and VanderWeele, 2012), sensitivity analyses for ranges of plausible associations between cross-world counterfactuals remain undeveloped. Further development of sensitivity analyses of both forms would be highly beneficial for practical use, and is fertile ground for future work. We hope that the work presented here will inspire deeper consideration and transparency regarding underlying identifying assumptions in the practice of mediation analysis.

2. PRELIMINARIES

By way of introduction, the directed acyclic graph (DAG) displayed in Fig. 1.(a) illustrates the simplest possible mediation setting, where A is defined to be the exposure taking either baseline value a^* or comparison value a, M is defined to be the (potential) mediator, and Y is defined to be the outcome. This DAG assumes randomization of the exposure, which for expositional

simplicity we maintain throughout. The graph also encodes no unobserved confounding of the effect of M on Y given A. The effect along the path $A \rightarrow Y$ on the diagram is generally referred to as direct with respect to M, and the effect along the path $A \rightarrow M \rightarrow Y$ on the diagram is generally referred to as indirect with respect to M.

Further elaboration of the specific type of direct and indirect effect under consideration necessitates counterfactual definitions. Let $Y(a)$ denote a subject's outcome if treatment A were set, possibly contrary to fact, to a. In the context of mediation, there will also be potential outcomes for the intermediate variable. Counterfactuals $M(a)$ and $Y(m, a)$ are defined similarly. In order to link these with the observed data, we adopt the standard set of consistency assumptions that

if $A = a$, then $M(a) = M$ with probability one,
if $A = a$ and $M = m$, then $Y(m, a) = Y$ with probability one, and
if $A = a$, then $Y(a) = Y$ with probability one.

In terms of counterfactuals, the randomization assumption encoded by the DAG in Fig. 1.(a) is $\{Y(a, m), M(a)\} \perp\!\!\!\perp A$ for all a and m; the assumption of no unobserved confounding of M given A is $Y(a, m) \perp\!\!\!\perp M(a) \mid A = a$ for all a and m. Finally, we will consider as well defined the nested counterfactual $Y\{a, M(a^*)\}$, i.e., the counterfactual outcome under an intervention which sets the exposure to the comparison value a, and the mediator to the value it would have taken under the conflicting baseline exposure value a^*.

We may now define the pure/natural direct effect and natural indirect effect (Robins and Greenland, 1992; Pearl, 2001), which form the following decomposition of the average causal effect:

$$
\begin{aligned}
& E\{Y(a)\} - E\{Y(a^*)\} \\
&= \overbrace{E[Y\{a, M(a)\}] - E[Y\{a^*, M(a^*)\}]}^{\text{total effect}} \\
&= \overbrace{E[Y\{a, M(a)\}] - E[Y\{a, M(a^*)\}]}^{\text{natural indirect effect}} + \overbrace{E[Y\{a, M(a^*)\}] - E[Y\{a^*, M(a^*)\})}^{\text{pure direct effect}}.
\end{aligned}
$$

The terms $E\{Y(a)\} = E\left[Y\left\{a, M(a)\right\}\right]$, for all a, are identified under randomization of A. The parameter $\gamma_0 \equiv E[Y\{a, M(a^*)\}]$ would be identified if one were to interpret the DAG in Fig. 1.(a) as a nonparametric structural equation model with independent errors (NPSEM-IE). Structural equations provide a nonparametric algebraic interpretation of this DAG corresponding to three equations, one for each variable in the graph. Each random variable on the graph is associated with a distinct, arbitrary function, denoted g, and a distinct random disturbance, denoted ε, 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 (i.e., its parents on the graph), and its corresponding random disturbance, as follows:

$$\begin{aligned} A &= g_A(\varepsilon_A) \\ M &= g_M(A, \varepsilon_M) \\ Y &= g_Y(A, M, \varepsilon_Y). \end{aligned}$$

Under particular interventions, these structural equations naturally encode dependencies of counterfactuals. Consider, for example, two interventions, one setting $A = a^*$, and another setting $A = a$ and $M = m$. The structural equations then become

$$\begin{aligned} A &= a^* & \qquad A &= a \\ M(a^*) &= g_M(a^*, \varepsilon_M) & \qquad M(a) &= m \\ Y(a^*) &= g_Y(a^*, M(a^*), \varepsilon_Y) & \qquad Y(a, m) &= g_Y(a, m, \varepsilon_Y). \end{aligned}$$

This formulation places no a priori restriction on the distribution of counterfactuals. The key assumption of the NPSEM-IE is that the random disturbances are mutually independent. This allows us to make independence statements regarding counterfactuals under various, possibly-conflicting interventions. In particular, this model implies that for all m, (i) $\{M(a), Y(a, m)\} \perp\!\!\!\perp A$, (ii) $Y(a, m) \perp\!\!\!\perp M(a) \mid A = a$, and (iii) $Y(a, m) \perp\!\!\!\perp M(a^*) \mid A = a$, which in turn suffice for identification of γ_0 (Pearl, 2001). Independence statements such as (iii) are known as cross-world counterfactual statements if a is not equal to a^*, due to their comparison of interventions that could never occur in the same world simultaneously. Independence condition (iii) can be seen to hold under the model by considering the NPSEM-IE under a specific intervention and noting

that the only source of randomness in $Y(a,m) = g_Y(a,m,\varepsilon_Y)$ is ε_Y and the only source of randomness in $M(a^*) = g_M(a^*,\varepsilon_M)$ is ε_M. Thus, the cross-world-counterfactual-independence statement follows directly from independence of exogenous disturbances. However, such an independence is neither experimentally verifiable nor enforceable (Robins and Richardson, 2010).

This issue has been discussed extensively (Robins and Richardson, 2010; Richardson and Robins, 2013), and in large part motivated the development of the single-world intervention graphs (SWIGs) of Richardson and Robins (2013). These causal graphs manage to elucidate this issue by graphically representing the counterfactuals themselves, allowing independence statements of counterfactuals to be read directly from the graph. Consider the SWIG in Fig. 1.(b). By d-separation, it is clear that (i) $Y(a,m) \perp\!\!\!\perp M(a)$ for all a and m, however no such statement can be made from the graph about $Y(a,m)$ and $M(a^*)$ when $a \neq a^*$. Under this SWIG, independence between $Y(a,m)$ and $M(a^*)$ is not assumed, and hence γ_0 is not point identified. Robins and Richardson (2010) provide the following bounds for its partial identification in the setting where M is binary and SWIG independence assumptions $M(a) \perp\!\!\!\perp A$ and $Y(a,m) \perp\!\!\!\perp \{M(a), A\}$ hold for all a and m:

$$\begin{aligned}
&\max\{0, \text{pr}(M=0 \mid A=a^*) + E(Y \mid M=0, A=a) - 1\} \\
&\quad + \max\{0, \text{pr}(M=1 \mid A=a^*) + E(Y \mid M=1, A=a) - 1\} \\
&\qquad\qquad \leq \gamma_0 \leq \\
&\min\{\text{pr}(M=0 \mid A=a^*), E(Y \mid M=0, A=a)\} \\
&\quad + \min\{\text{pr}(M=1 \mid A=a^*), E(Y \mid M=1, A=a)\}.
\end{aligned}$$

In Section 2, we extend this result to the setting of a polytomous M.

As previously mentioned, another often-overlooked condition required for identification of γ_0 is that there is no confounder of the mediator's effect on the outcome that is affected by the exposure. Such a confounder is present in the setting illustrated in the DAG in Fig. 2.(a). Generally, even under an NPSEM-IE interpretation of this DAG, γ_0 will not be identified in this setting. This is readily seen by considering the following representation under this model given

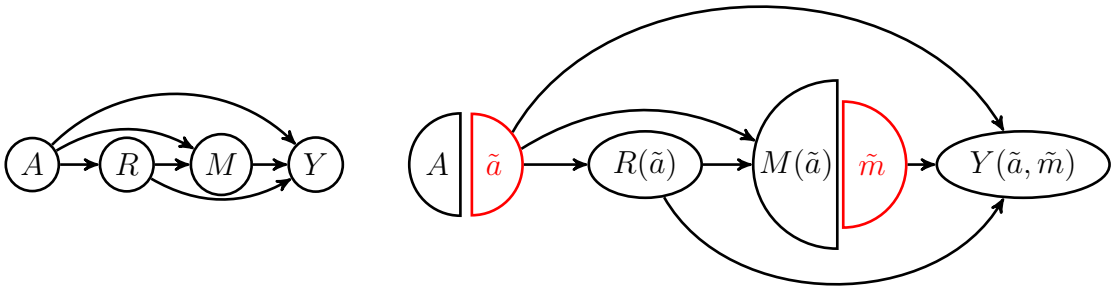

Figure 2: (a) A mediation directed acyclic graph in which R is an exposure-induced confounder. The nodes represent random variables, and the arrows represent possible causal effects of one random variable on another. (b) The single-world intervention graph in the setting of (a) that has been intervened on to set A to $\tilde{a} \in \{a, a^*\}$ and M to $\tilde{m}$. The black nodes represent random variables under this intervention, the red nodes represent the level an intervened random variable takes under this intervention, and the arrows represent possible causal effects of one variable under this intervention on another.

by Robins and Richardson (2010):

$$\gamma_0 = \sum_{r,r^*} E\left\{E(Y \mid M, R = r, A = a) \mid R = r^*, A = a^*\right\} \text{pr}\left\{R(a) = r, R(a^*) = r^*\right\}. \quad (1)$$

Clearly the joint probability term can never be identified from observed data, since we will never be able to observe $R(a)$ and $R(a^*)$ for the same individual.

A few conditions for identification have been proposed. Robins and Richardson (2010) give two. The first is that $R(a) \perp\!\!\!\perp R(a^*)$, in which case the troublesome term in (1) will factor, giving

$$\gamma_0 = \sum_{r^*,r} E\left\{E(Y \mid M, R = r, A = a) \mid R = r^*, A = a^*\right\} \text{pr}(R = r^* \mid A = a^*) \times \text{pr}(R = r \mid A = a).$$

It seems biologically unlikely, however, that in a scenario in which A affects R, the counterfactual R under $A = a$ would not be predictive of the counterfactual R under $A = a^*$. The other condition is that the counterfactual outcome under one exposure value is a deterministic function

of the counterfactual for the other treatment, i.e., $R(a) = g\{R(a^*)\}$. In this case,

$$\gamma_0 = \sum_{r^*,r} E\left\{E(Y \mid M, R = r, A = a) \mid R = r^*, A = a^*\right\} \text{pr}(R = r^* \mid A = a^*) I\{r = g(r^*)\}.$$

The above assumption is implied by rank preservation (Robins and Richardson, 2010), which is unlikely to hold in social and health sciences as it rules out individual-level effect heterogeneity (Tchetgen Tchetgen and VanderWeele, 2014). As none of these conditions are experimentally verifiable, the authors themselves "do not advocate blithely adopting such assumptions in order to preserve identification of the PDE in [this setting]" (Robins and Richardson, 2010).

Tchetgen Tchetgen and VanderWeele (2014) give two testable conditions for identification of γ_0 when R is present. The first is of A–R monotonicity, i.e., for Bernoulli R, $R(a) \geq R(a^*)$. If R is a vector of Bernoulli random variables whose structural equations have independent errors, and if monotonicity holds for each element,

$$\gamma_0 = \sum_{r,r^*} E\left\{E(Y \mid M, R = r, A = a) \mid R = r^*, A = a^*\right\} \prod_{j=1}^{k} f_j(r_j, r_j^*, a, a^*)$$

where

$$f_j(r_j, r_j^*, a, a^*) = \begin{cases} \text{pr}(R_j = 1 \mid A = a^*) & \text{if } r_j^* = r_j = 1, \\ \text{pr}(R_j = 1 \mid A = a) - \text{pr}(R_j = 1 \mid A = a^*) & \text{if } r_j^* = 0 \text{ and } r_j = 1, \\ 0 & \text{if } r_j^* = 1 \text{ and } r_j = 0, \\ \text{pr}(R_j = 0 \mid A = a) & \text{if } r_j^* = r_j = 0. \end{cases}$$

Their second condition is no M–R additive mean interaction, i.e.,

$$E(Y \mid m, r, a) - E(Y \mid m^*, r, a) - E(Y \mid m, r^*, a) + E(Y \mid m^*, r^*, a) = 0,$$

for all levels m and m^* of M and r and r^* of R. For discrete M and R, this yields

$$\begin{aligned}\gamma_0 = &\sum_m \{E(Y \mid m, r^*, a) - E(Y \mid m^*, r^*, a)\} \operatorname{pr}(M = m \mid A = a^*) \\ &+ \sum_r \{E(Y \mid m^*, r, a) - E(Y \mid m^*, r^*, a)\} \operatorname{pr}(R = r \mid A = a) \\ &+ E(Y \mid m^*, r^*, a).\end{aligned}$$

Eschewing the cross-world-counterfactual assumptions of the NPSEM-IE , Tchetgen Tchetgen and Phiri (2014) extend the bounds of Robins and Richardson (2010) to allow for the presence of an exposure-induced confounder when the mediator is binary:

$$\begin{aligned}&\max\left\{0, \operatorname{pr}(M = 0 \mid A = a^*) + \sum_r E(Y \mid M = 0, R = r, A = a)\operatorname{pr}(R = r \mid A = a) - 1\right\} \\ &+ \max\left\{0, \operatorname{pr}(M = 1 \mid A = a^*) + \sum_r E(Y \mid M = 1, R = r, A = a)\operatorname{pr}(R = r \mid A = a) - 1\right\} \\ &\qquad\qquad \leq \gamma_0 \leq \\ &\min\left\{\operatorname{pr}(M = 0 \mid A = a^*), \sum_r E(Y \mid M = 0, R = r, A = a)\operatorname{pr}(R = r \mid A = a)\right\} \\ &+ \min\left\{\operatorname{pr}(M = 1 \mid A = a^*), \sum_r E(Y \mid M = 1, R = r, A = a)\operatorname{pr}(R = r \mid A = a)\right\}.\end{aligned}$$

We extend these bounds as well to allow for polytomous M in Section 3. Additionally, we construct bounds for γ_0 under an NPSEM-IE that account for a discrete exposure-induced confounder, but require no further assumption.

3. NEW PARTIAL IDENTIFICATION RESULTS

We begin by extending the bounds of Robins and Richardson (2010) and Tchetgen Tchetgen and Phiri (2014) to settings with discrete mediator and outcome. Proofs can be found in the Appendix.

Theorem 1. *Under the* SWIG *in either Fig. 1.(b) or Fig. 2.(b) with discrete M and Y and*

arbitrary R,

$$\sum_{m,y} y \left(\max\left[0, \text{pr}\{M(a^*) = m\} + \text{pr}\{Y(a,m) = y\} - 1\right] I(y > 0)\right.$$
$$\left. + \min\left[\text{pr}\{M(a^*) = m\}, \text{pr}\{Y(a,m) = y\}\right] I(y < 0)\right)$$
$$\leq \gamma_0 \leq$$
$$\sum_{m,y} y \left(\max\left[0, \text{pr}\{M(a^*) = m\} + \text{pr}\{Y(a,m) = y\} - 1\right] I(y < 0)\right.$$
$$\left. + \min\left[\text{pr}\{M(a^*) = m\}, \text{pr}\{Y(a,m) = y\}\right] I(y > 0)\right).$$

The upper and lower bounds coincide when $Y(a,m)$ or $M(a^*)$ is degenerate, which follows from the properties of joint probability mass functions. The upper bound is achieved only if $Y(a,m)$ and $M(a^*)$ are comonotone for each m, i.e., if $F_{Y(a,m),M(a^*)}(y,m) = \min\left[F_{Y(a,m)}(y), F_{M(a^*)}(m)\right]$ for each m, where $F_X(\cdot)$ denotes the joint (or marginal) cumulative distribution function of the random vector (or scalar) X. The lower bound is achieved only if they are countermonotone for each m, i.e., if $F_{Y(a,m),M(a^*)}(y,m) = \max\left\{0, F_{Y(a,m)}(y) + F_{M(a^*)}(m) - 1\right\}$ for each m. A straightforward application of the g-formula under the DAGs in Fig. 1 and 2 yields the following corollaries:

Corollary 1. *For polytomous M and Y, γ_0 is partially identified under the* SWIG *in Fig. 1.(b) by the bounds in Theorem 1 with* $\text{pr}\{M(a^*) = m\} = \text{pr}(M = m \mid a^*)$ *and* $\text{pr}\{Y(a,m) = y\} = \text{pr}(Y = y \mid m, a)$. *It is partially identified under the* SWIG *in Fig. 2.(b) by the same bounds, but with* $\text{pr}\{M(a^*) = m\} = \text{pr}(M = m \mid a^*)$ *and* $\text{pr}\{Y(a,m) = y\} = \sum_r \text{pr}(Y = y \mid m, r, a)\text{pr}(R = r \mid a)$.

The second part of the corollary continues to hold even if there were a hidden common cause of R and Y as in Fig. 3, since the same g-formula applies in this setting. Whereas the previous results invoked no cross-world-counterfactual independences under the SWIG interpretation of the DAG in Fig. 2.(a), sharper bounds are available under Pearl's NPSEM-IE interpretation of these DAGs, as derived in the following result.

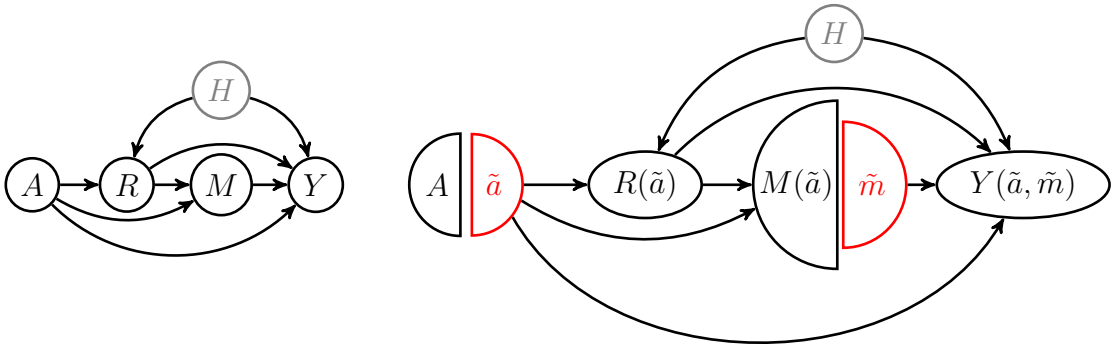

Figure 3: (a) A mediation directed acyclic graph in which an unobserved variable H affects R, an exposure-induced confounder, and Y. The black nodes represent observed random variables, and the arrows represent possible causal effects of one random variable on another. (b) The single-world intervention graph in the setting of (a) that has been intervened on to set A to $\tilde{a} \in \{a, a^*\}$ and M to $\tilde{m}$. The black nodes represent random variables under this intervention, the red nodes represent the level an intervened random variable takes under this intervention, and the arrows represent possible causal effects of one variable under this intervention on another. In each panel, the gray node represents a hidden random variable

Theorem 2. *For discrete R taking values in $\{1, \ldots, p\}$, let B be the $p^2 \times (p-1)^2$ matrix*

$$\begin{bmatrix}
I_{p-1} & 0_{(p-1)\times(p-1)} & \cdots & 0_{(p-1)\times(p-1)} & 0_{(p-1)\times(p-1)} \\
-1_{p-1}^T & 0_{p-1}^T & \cdots & 0_{p-1}^T & 0_{p-1}^T \\
0_{(p-1)\times(p-1)} & I_{p-1} & \cdots & 0_{(p-1)\times(p-1)} & 0_{(p-1)\times(p-1)} \\
0_{p-1}^T & -1_{p-1}^T & \cdots & 0_{p-1}^T & 0_{p-1}^T \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0_{(p-1)\times(p-1)} & 0_{(p-1)\times(p-1)} & \cdots & I_{p-1} & 0_{(p-1)\times(p-1)} \\
0_{p-1}^T & 0_{p-1}^T & \cdots & -1_{p-1}^T & 0_{p-1}^T \\
0_{(p-1)\times(p-1)} & 0_{(p-1)\times(p-1)} & \cdots & 0_{(p-1)\times(p-1)} & I_{p-1} \\
0_{p-1}^T & 0_{p-1}^T & \cdots & 0_{p-1}^T & -1_{p-1}^T \\
-I_{p-1} & -I_{p-1} & \cdots & -I_{p-1} & -I_{p-1} \\
 & & 1_{(p-1)^2}^T & &
\end{bmatrix},$$

d be the p^2-dimensional vector

$$\begin{bmatrix} 0_{p-1} \\ \text{pr}\left(R=1 \mid A=a\right) \\ 0_{p-1} \\ \text{pr}\left(R=2 \mid A=a\right) \\ \vdots \\ 0_{p-1} \\ \text{pr}\left(R=p-1 \mid A=a\right) \\ \text{pr}\left(R=1 \mid A=a^*\right) \\ \text{pr}\left(R=2 \mid A=a^*\right) \\ \vdots \\ \text{pr}\left(R=p-1 \mid A=a^*\right) \\ \text{pr}\left(R=p \mid A=a\right)+\text{pr}\left(R=p \mid A=a^*\right)-1 \end{bmatrix},$$

and x be the vectorization of the matrix $\left[E\left\{E(Y \mid M, R=r, A=a) \mid R=r^*, A=a^*\right\}\right]_{r,r^*}$. Under a NPSEM-IE *corresponding to the* DAG *in Fig. 2.(a) where M and Y can be either continuous or discrete, γ_0 is partially identified by $\left[x^T(B\delta_L+d), x^T(B\delta_U+d)\right]$, where δ_L and δ_U are the minimizing and maximizing solutions respectively to the linear programming problem with objective function $x^TB\delta$ subject to the constraints*

$$\begin{bmatrix} I_{(p-1)^2} \\ -I_{(p-1)^2} \end{bmatrix} \delta \leq \begin{bmatrix} \min\{\text{pr}(R=1 \mid A=a), \text{pr}(R=1 \mid A=a^*)\} \\ \min\{\text{pr}(R=1 \mid A=a), \text{pr}(R=2 \mid A=a^*)\} \\ \vdots \\ \min\{\text{pr}(R=p \mid A=a), \text{pr}(R=p-1 \mid A=a^*)\} \\ \min\{\text{pr}(R=p \mid A=a), \text{pr}(R=p \mid A=a^*)\} \\ \min\{0, 1-\text{pr}(R=1 \mid A=a)-\text{pr}(R=1 \mid A=a^*)\} \\ \min\{0, 1-\text{pr}(R=1 \mid A=a)-\text{pr}(R=2 \mid A=a^*)\} \\ \vdots \\ \min\{0, 1-\text{pr}(R=p \mid A=a)-\text{pr}(R=p-1 \mid A=a^*)\} \\ \min\{0, 1-\text{pr}(R=p \mid A=a)-\text{pr}(R=p \mid A=a^*)\} \end{bmatrix}$$

and $\delta \geq 0$.

Similar to the previous result, these bounds coincide if either $R(a)$ or $R(a^*)$ is degenerate. The upper bound is achieved when $R(a)$ and $R(a^*)$ are comonotone; the lower bound is achieved when they are countermonotone. While these bounds are not available in closed form, they can be readily solved using standard software, such as with the lp_solve function, which uses the revised simplex method and is accessible from a number of languages, including R, MATLAB, Python, and C. While the method used by this software is not guaranteed to converge at a polynomial rate (Klee and Minty, 1970), it is quite efficient in most cases (Schrijver, 1998). The following corollary shows that these bounds reduce to a closed form when R is binary.

Corollary 2. *Under a* NPSEM-IE *corresponding to the* DAG *in Fig. 2.(a) with binary* R*,*

$$\min_{\pi_{11}\in\Pi} \sum_{r,r^*} E\left\{E(Y \mid M, R=r, A=a) \mid R=r^*, A=a^*\right\} h(r, r^*, \pi_{11})$$

$$\leq \gamma_0 \leq$$

$$\max_{\pi_{11}\in\Pi} \sum_{r,r^*} E\left\{E(Y \mid M, R=r, A=a) \mid R=r^*, A=a^*\right\} h(r, r^*, \pi_{11})$$

where Π *is the set*

$$\{\max\{0, \mathrm{pr}(R=1 \mid A=a) + \mathrm{pr}(R=1 \mid A=a^*) - 1\},$$
$$\min\{\mathrm{pr}(R=1 \mid A=a), \mathrm{pr}(R=1 \mid A=a^*)\}\}$$

and

$$h(r, r^*, \pi_{11}) = \begin{cases} \pi_{11} & \text{if } r^*=r=1, \\ \mathrm{pr}(R=1 \mid A=a) - \pi_{11} & \text{if } r^*=0 \text{ and } r=1, \\ \mathrm{pr}(R=1 \mid A=a^*) - \pi_{11} & \text{if } r^*=1 \text{ and } r=0, \\ 1 - \mathrm{pr}(R=1 \mid A=a) - \mathrm{pr}(R=1 \mid A=a^*) + \pi_{11} & \text{if } r^*=r=0. \end{cases}$$

Under $A-R$ monotonicity with binary R, the identifying functional given by Tchetgen Tchetgen and VanderWeele (2014) is recovered at the upper bound in Corollary 2. All results given

here can be extended to settings with observed pre-exposure confounders, which we denote C. In Corollary 1, one must first perform conditional inference given C, then subsequently average over the conditional bounds. This is in fact valid due to Jensen's inequality, because the constraints on the marginal joint probabilities are already implied by the constraints enforced on the conditional joint distributions, so no further constraints need be considered. However, Jensen's inequality does not apply in the case of Theorem 2, so controlling for C requires estimating two pairs of candidate bounds and selecting the larger of the lower bounds and the smaller of the upper bounds. When p is of moderate size, δ can be solved for each covariate pattern of C, i.e., without modeling the dependence of the cross-world-counterfactual joint distribution on C. Averaging the resulting conditional bounds gives the first pair of bounds. The second pair results from replacing each probability in the theorem with an average over the probabilities conditional on C and doing the same with x.

4. APPLICATION TO HARVARD PEPFAR DATA SET

We now consider an application to a data set collected by the Harvard President's Emergency Plan for AIDS Relief (PEPFAR) program in Nigeria. The data set consists of previously antiretroviral therapy (ART)-naïve, HIV-1 infected adult patients who began ART in the program and were followed at least one year following initiation. Patients without reliable viral load data at two of the hospitals were excluded. Only complete cases initially prescribed to either TDF+3TC/FTC+NVP or AZT+3TC+NVP[1] were considered for this analysis. Thus, the data set we consider consists of 6627 patients, 1919 of whom were prescribed to TDF+3TC/FTC+NVP, and the remaining 4708 prescribed to AZT+3TC+NVP.

There has accumulated evidence of a differential effect on virologic failure between these two first-line antiretroviral treatment regimens (Tang et al., 2012). Virologic failure is defined by the World Health Organization as repeat viral load above 1000 copies/mL. We base this on measurements at 12 and 18 months of ART duration in our analysis.

A natural question of scientific interest is what role adherence plays in mediating this differential effect. We are primarily interested in learning about the scientific mechanism of this effect on the individual level. The natural indirect effect best captures this mechanism in that it captures

[1] 3TC=lamivudine, AZT=zidovudine, FTC=emtricitabine, NVP=nevirapine, TDF=tenofovir

an isolated effect difference mediated by adherence by, in a sense, deactivating effect differences along all other possible causal pathways. We specifically examine the effect through adherence over the second six months since treatment assignment, i.e., the six months prior to the first viral load measurement. Identification is complicated by the presence of treatment toxicity, which clearly affects adherence directly, and has the potential to modify the effect of the treatment assignment on virologic failure. Thus, toxicity measured at six months after treatment assignment is an exposure-induced confounder of the effect of the mediator on the outcome. Further, toxicity and virologic failure are likely to be rendered dependent by unobserved underlying biological common causes as in Fig. 3, where H represents these hidden biological mechanisms. Because we define the mediator to be adherence over the second six months, adherence over the first six months is also an exposure-induced confounder along with toxicity, and must be accounted for. Had we defined the mediator to be adherence over the full year, measurement of the mediator and toxicity would have overlapped, violating the principle of temporal ordering.

Let C denote the vector consisting of baseline covariates sex, age, marital status, WHO stage, hepatitis C virus, hepatitis B virus, CD4+ cell count, and viral load. Let A be an indicator of ART assignment taking levels a^* for TDF+3TC/FTC+NVP and a for AZT+3TC+NVP; R be a vector of two indicator variables, one of the presence of any lab toxicity, and one of adherence exceeding 95%, both over the first six months following initiation of therapy; M be an indicator of adherence exceeding 95% over the subsequent six months; and Y be an indicator of virologic failure at one year, i.e., repeat viral load above 1000 copies/mL at one year and at 18 months.

Here we estimate the natural indirect effect of A on Y through M, as defined above, on the risk difference scale using the various sets of identifying assumptions given above. Throughout, inference is performed using maximum likelihood for point estimation and a weighted bootstrap (Rao and Zhao, 1992; van der Vaart and Wellner, 1996) for confidence intervals, which appropriately accounts for the rare outcome. The results are summarized in Fig. 4. It is immediately apparent that inference is sensitive to which identifying assumptions are made. Consider an investigator who might be willing to rely on cross-world-counterfactual independences. If she decides to ignore the presence of toxicity, she might likely conclude that there is a very small, yet significant negative indirect effect. Conversely, were she to make the no M–R interaction assumption, she would find a significant positive indirect effect with considerable uncertainty.

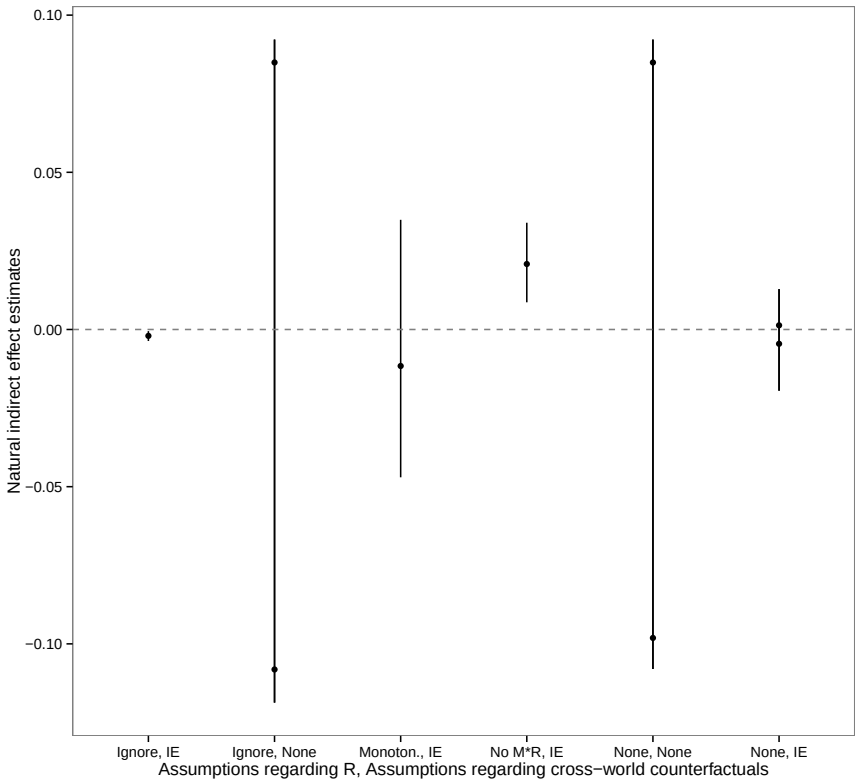

Figure 4: A plot showing the estimated natural indirect effect of ART assignment on virologic failure with respect to adherence under the various assumptions. The assumptions vary across the horizontal axis, with the first part of the label indicating the assumption regarding the exposure-induced confounder, R, and the second part indicating the assumption regarding cross-world counterfactuals. For the assumptions regarding R, "Ignore" means that the presence of R is ignored altogether, "Monoton." means the A–R monotonicity assumption in Section 1, "No M*R" means the no M–R interaction assumption in Section 1, and "None" means that R was accounted for without additional assumptions. For the assumptions regarding cross-world counterfactuals, "IE" means a NPSEM-IE was assumed, and "None" means no cross-world-counterfactuals independences were assumed. When the assumptions give partial identification, the two dots represent the point estimates of the upper and lower bound for the natural indirect effect, and the vertical bar represents the bootstrap 95% confidence interval for the interval. When the assumptions give full identification, the single dot represents the point estimate of the natural indirect effect, and the vertical bar represents its bootstrap 95% confidence interval.

In fact, an empirical test of this assumption reveals that it is unlikely to apply. Likewise, the data suggest that the required assumption of independent errors of the components of R is also unlikely to hold. Nonetheless, we present both results for the sake of comparison. Results are fairly imprecise under monotonicity, and do not show a significant effect.

Another investigator unwilling to impose cross-world-counterfactual-independence assumptions is left with little to say as the bounds are wide, and include the null hypothesis of no NIE, regardless of how toxicity is handled. Interestingly, the bounds that result from making no assumptions about the joint distribution of the cross-world R counterfactuals are narrower than the bounds that result from ignoring R. That is, the bounds themselves appear narrower; the variances of the interval estimates appear to be comparable. This is because even though we do not impose any restrictions on the distribution of R or its counterfactuals a priori, observing R is clearly informative. The bounds accounting for R have the added advantage of being the only identifying formula that remains valid when toxicity and virologic suppression are affected by an unobserved common cause, as in Fig. 3.

Finally, incorporating R results in narrower interval estimates than not imposing the NPSEM-IE, even if R were ignored. Thus, cross-world-counterfactual-independences appear to have stronger empirical implications in the current analysis than assumptions regarding exposure-induced confounders. The general trend in these results is that little is gained in terms of precision by assumptions regarding R. In fact, the confidence interval for the bounds resulting from the independent errors assumption and no assumption regarding R is narrower than the confidence interval for the estimate that results from assuming monotonicity, despite the fact that the NIE is point-identified in the latter case. The naïve assumption that R is not a confounder is the only assumption about R under which precision is gained.

APPENDIX

Proofs of theorems

Proof of Theorem 1. For each level m and y, define $\pi_1(m,y) = \text{pr}\{Y(a,m) = y\}$ and $\pi_2(m) = \text{pr}\{M(a^*) = m\}$. There exist $U_1(m,y), U_2(m) \sim \mathcal{U}(0,1)$ such that $I\{Y(a,m) = y\} = I\{U_1(m,y) \leq \pi_1(m,y)\}$ and $I\{M(a^*) = m\} = I\{U_2(m) \leq \pi_2(m)\}$. The joint distribution

$F_{U_1(m,y),U_2(m)}$, then, is a bivariate copula, for which Fréchet–Hoeffding sharp bounds exist. Applying these to $\text{pr}\left\{Y(a,m)=y, M(a^*)=m\right\} = F_{U_1(m,y),U_2(m)}\left\{\pi_1(m,y),\pi_2(m)\right\}$, we have

$$\begin{aligned}&\max\left[0, \text{pr}\{M(a^*)=m\} + \text{pr}\{Y(a,m)=y\} - 1\right]\\ &\leq \text{pr}\left\{Y(a,m)=y, M(a^*)=m\right\} \leq\\ &\min\left[\text{pr}\{M(a^*)=m\}, \text{pr}\{Y(a,m)=y\}\right].\end{aligned}$$

Applying these bounds to each summand in

$$E[Y\{a, M(a^*)\}] = \sum_{m,y} y\text{pr}\{Y(a,m)=y, M(a^*)=m\}$$

yields the result. □

Proof of Theorem 2. Let $\pi_{r,r^*} = \text{pr}\left\{R(a)=r, R(a^*)=r^*\right\}$, π be the vectorization of the matrix $[\pi_{r,r^*}]$, and δ be the vectorization of the matrix $[\pi_{r,r^*}]_{-p,-p}$, i.e., the vectorization of the matrix π with row p and column p removed. Equation (1) can now be expressed as $\gamma_0 = x^T\pi$, which is identified in x, but not π. Conditional on the marginal probabilities, which are identified, the joint probabilities have $(p-1)^2$ degrees of freedom, and can be expressed as $\pi = B\delta + d$. Since $x^TB\delta$ is linear in δ and each element of δ is constrained by

$$\begin{aligned}&\max\left\{0, \text{pr}(R=r \mid A=a) + \text{pr}(R=r^* \mid A=a^*) - 1\right\}\\ &\leq \pi_{r,r^*} \leq\\ &\min\left\{\text{pr}(R=r \mid A=a), \text{pr}(R=r^* \mid A=a^*)\right\},\end{aligned}$$

the proposed linear programming problem will yield the δ that optimizes $x^TB\delta$, and hence $x^T(B\delta+d)$. Thus, γ_0 will be bounded by $x^T(B\delta+d)$ evaluated at the minimizing and maximizing linear programming solutions δ_L and δ_U. □

REFERENCES

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in Medicine*, 27(8):1282–1304.

Albert, J. M. (2012). Mediation analysis for nonlinear models with confounding. *Epidemiology (Cambridge, Mass.)*, 23(6):879.

Albert, J. M. and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, 67(3):1028–1038.

Albert, J. M. and Wang, W. (2015). Sensitivity analyses for parametric causal mediation effect estimation. *Biostatistics*, 16(2):339–351.

Avin, C., Shpitser, I., and Pearl, J. (2005). Identifiability of path-specific effects. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 357–363.

Balke, A. A. and Pearl, J. (1997). Probabilistic counterfactuals: semantics, computation, and applications. Technical report, DTIC Document.

Cai, Z., Kuroki, M., Pearl, J., and Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):815–836.

Goetgeluk, S., Vansteelandt, S., and Goetghebeur, E. (2008). Estimation of controlled direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):1049–1066.

Hsu, J. Y., Kennedy, E. H., Roy, J. A., Stephens-Shields, A. J., Small, D. S., and Joffe, M. M. (2015). Surrogate markers for time-varying treatments and outcomes. *Clinical Trials*, page 1740774515583500.

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

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71.

Kaufman, S., Kaufman, J. S., MacLehose, R. F., Greenland, S., and Poole, C. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine*, 24(11):1683–1702.

Klee, V. and Minty, G. J. (1970). How good is the simplex algorithm. Technical report, DTIC Document.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

Petersen, M. L., Sinisi, S. E., and van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, 17(3):276–284.

Rao, C. R. and Zhao, L. (1992). Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 323–331.

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

Robins, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113:159.

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

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. *Highly Structured Stochastic Systems*, pages 70–81.

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

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, pages 103–158.

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

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

Schrijver, A. (1998). *Theory of linear and integer programming*. John Wiley & Sons.

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

Sjölander, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine*, 28(4):558–571.

Splawa-Neyman, J., Dabrowska, D., Speed, T., et al. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472.

Taguri, M. and Chiba, Y. (2015). A principal stratification approach for evaluating natural direct and indirect effects in the presence of treatment-induced intermediate confounding. *Statistics in Medicine*, 34(1):131–144.

Tang, M. W., Kanki, P. J., and Shafer, R. W. (2012). A review of the virological efficacy of the 4 World Health Organization–recommended tenofovir-containing regimens for initial HIV therapy. *Clinical Infectious Diseases*, 54(6):862–875.

Taylor, J. M., Wang, Y., and Thiébaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 61(4):1102–1111.

Tchetgen Tchetgen, E. J. (2011). On causal mediation analysis with a survival outcome. *The International Journal of Biostatistics*, 7(1):1–38.

Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine*, 32(26):4567–4580.

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

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

Tchetgen Tchetgen, E. J. and Shpitser, I. (2014). Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864.

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

Ten Have, T. R., Joffe, M. M., Lynch, K. G., Brown, G. K., Maisto, S. A., and Beck, A. T. (2007). Causal mediation analyses with rank preserving models. *Biometrics*, 63(3):926–934.

van der Laan, M. J. and Petersen, M. L. (2008). Direct effect models. *The International Journal of Biostatistics*, 4(1):1–27.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

VanderWeele, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26.

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology (Cambridge, Mass.)*, 22(4):582.

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

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172(12):1339–1348.

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

Wang, W. and Albert, J. M. (2012). Estimation of mediation effects for zero-inflated regression models. *Statistics in Medicine*, 31(26):3118–3132.

Wang, W., Nelson, S., and Albert, J. M. (2013). Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Statistics in Medicine*, 32(24):4211–4228.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by âĂIJdeathâĂİ. *Journal of Educational and Behavioral Statistics*, 28(4):353–368.