The Annals of Statistics
2012, Vol. 40, No. 3, 1816–1845
DOI: 10.1214/12-AOS990
© Institute of Mathematical Statistics, 2012

SEMIPARAMETRIC THEORY FOR CAUSAL MEDIATION ANALYSIS: EFFICIENCY BOUNDS, MULTIPLE ROBUSTNESS AND SENSITIVITY ANALYSIS

BY ERIC J. TCHETGEN TCHETGEN[1] AND ILYA SHPITSER

Harvard School of Public Health

While estimation of the marginal (total) causal effect of a point exposure on an outcome is arguably the most common objective of experimental and observational studies in the health and social sciences, in recent years, investigators have also become increasingly interested in mediation analysis. Specifically, upon evaluating the total effect of the exposure, investigators routinely wish to make inferences about the direct or indirect pathways of the effect of the exposure, through a mediator variable or not, that occurs subsequently to the exposure and prior to the outcome. Although powerful semiparametric methodologies have been developed to analyze observational studies that produce double robust and highly efficient estimates of the marginal total causal effect, similar methods for mediation analysis are currently lacking. Thus, this paper develops a general semiparametric framework for obtaining inferences about so-called marginal natural direct and indirect causal effects, while appropriately accounting for a large number of pre-exposure confounding factors for the exposure and the mediator variables. Our analytic framework is particularly appealing, because it gives new insights on issues of efficiency and robustness in the context of mediation analysis. In particular, we propose new multiply robust locally efficient estimators of the marginal natural indirect and direct causal effects, and develop a novel double robust sensitivity analysis framework for the assumption of ignorability of the mediator variable.

1. Introduction. The evaluation of the total causal effect of a given point exposure, treatment or intervention on an outcome of interest is arguably the most common objective of experimental and observational studies in the fields of epidemiology, biostatistics and in the social sciences. However, in recent years, investigators in these various fields have become increasingly interested in making inferences about the direct or indirect pathways of the exposure effect, through a mediator variable or not, that occurs subsequently to the exposure and prior to the outcome. Recently, the counterfactual language of causal inference has proven particularly useful for formalizing mediation analysis. Indeed, causal inference

Received March 2011; revised March 2012.
[1]Supported by NIH Grant R21ES019712.
MSC2010 subject classifications. 62G05.
Key words and phrases. Natural direct effects, natural indirect effects, double robust, mediation analysis, local efficiency.

offers a formal mathematical framework for defining varieties of direct and indirect effects, and for establishing necessary and sufficient identifying conditions of these effects. A notable contribution of causal inference to the literature on mediation analysis is the key distinction drawn between so-called controlled direct effects versus natural direct effects. In words, the controlled direct effect refers to the exposure effect that arises upon intervening to set the mediator to a fixed level that may differ from its actual observed value [Robins and Greenland (1992), Pearl (2001), Robins (2003)]. In contrast, the natural (also known as pure) direct effect captures the effect of the exposure when one intervenes to set the mediator to the (random) level it would have been in the absence of exposure [Robins and Greenland (1992), Pearl (2001)]. As noted by Pearl (2001), controlled direct and indirect effects are particularly relevant for policy making, whereas natural direct and indirect effects are more useful for understanding the underlying mechanism by which the exposure operates. In fact, natural direct and indirect effects combine to produce the exposure total effect.

To formally define natural direct and indirect effects first requires defining counterfactuals. We assume that for each level of a binary exposure E, and of a mediator variable M, there exist a counterfactual variable $Y_{e,m}$ corresponding to the outcome Y had possibly contrary to fact the exposure and mediator variables taken the value (e, m). Similarly, for $E = e$, we assume there exists a counterfactual variable M_e corresponding to the mediator variable had possibly contrary to fact the exposure variable taken the value e. The current paper concerns the decomposition of the total effect of E on Y, in terms of natural direct and natural indirect effects, which, expressed on the mean difference scale, is given by

$$\begin{aligned}\overbrace{\mathbb{E}(Y_{e=1} - Y_{e=0})}^{\text{total effect}} &= \mathbb{E}(Y_{e=1,M_{e=1}} - Y_{e=0,M_{e=0}}) \\ &= \overbrace{\mathbb{E}(Y_{e=1,M_{e=1}} - Y_{e=1,M_{e=0}})}^{\text{natural indirect effect}} + \overbrace{\mathbb{E}(Y_{e=1,M_{e=0}} - Y_{e=0,M_{e=0}})}^{\text{natural direct effect}},\end{aligned} \tag{1}$$

where $\mathbb{E}$ stands for expectation.

In an effort to account for confounding bias when estimating causal effects, such as the average total effect (1) from nonexperimental data, investigators routinely collect and adjust for in data analysis, a large number of confounding factors. Because of the curse of dimensionality, nonparametric methods of estimation are typically not practical in such settings, and one usually resorts to one of two dimension-reduction strategies; either one relies on a model for the outcome given exposure and counfounders, or alternately one relies on a model for the exposure, that is, the propensity score. Recently, powerful semiparametric methods have been developed to analyze observational studies that produce so-called double robust and highly efficient estimates of the exposure total causal effect [Robins (2000), Scharfstein, Rotnitzky and Robins (1999), Bang and Robins (2005), Tsiatis (2006)] and similar methods have also been developed to estimate controlled direct

effects [Goetgeluk, Vansteelandt and Goetghebeur (2008)]. An important advantage of a double robust method is that it carefully combines both of the aforementioned dimension reduction strategies for confounding adjustment, to produce an estimator of the causal effect that remains consistent and asymptotically normal, provided at least one of the two strategies is correct, without necessarily knowing which strategy is indeed correct [van der Laan and Robins (2003)]. Unfortunately, similar methods for making semiparametric inferences about marginal natural direct and indirect effects are currently lacking. Thus, this paper develops a general semiparametric framework for obtaining inferences about marginal natural direct and indirect effects on the mean of an outcome, while appropriately accounting for a large number of confounding factors for the exposure and the mediator variables.

Our semiparametric framework is particularly appealing, as it gives new insight on issues of efficiency and robustness in the context of mediation analysis. Specifically, in Section 2, we adopt the sequential ignorability assumption of Imai, Keele and Tingley (2010) under which, in conjunction with the standard consistency and positivity assumptions, we derive the efficient influence function and thus obtain the semiparametric efficiency bound for the natural direct and natural indirect marginal mean causal effects, in the nonparametric model $\mathcal{M}_{\text{nonpar}}$ in which the observed data likelihood is left unrestricted. We further show that in order to conduct mediation inferences in $\mathcal{M}_{\text{nonpar}}$, one must estimate at least a subset of the following quantities:

(i) the conditional expectation of the outcome given the mediator, exposure and confounding factors;
(ii) the density of the mediator given the exposure and the confounders;
(iii) the density of the exposure given the confounders.

Ideally, to minimize the possibility of modeling bias, one may wish to estimate each of these quantities nonparametrically; however, as previously argued, when as we assume throughout, we wish to account for numerous confounders, such nonparametric estimates will likely perform poorly in finite samples. Thus, in Section 2.3 we develop an alternative multiply robust strategy. To do so, we propose to model (i), (ii) and (iii) parametrically (or semiparametrically), but rather than obtaining mediation inferences that rely on the correct specification of a specific subset of these models, instead we carefully combine these three models to produce estimators of the marginal mean direct and indirect effects that remain consistent and asymptotically normal (CAN) in a union model, where at least one but not necessarily all of the following conditions hold:

(a) the parametric or semi-parametric models for the conditional expectation of the outcome (i) and for the conditional density of the mediator (ii) are correctly specified;

(b) the parametric or semiparametric models for the conditional expectation of the outcome (i) and for the conditional density of the exposure (iii) are correctly specified;

(c) the parametric or semiparametric models for the conditional densities of the exposure and the mediator (ii) and (iii) are correctly specified.

Accordingly, we define submodels $\mathcal{M}_a$, $\mathcal{M}_b$ and $\mathcal{M}_c$ of $\mathcal{M}_{\text{nonpar}}$ corresponding to models (a), (b) and (c) respectively. Thus, the proposed approach is triply robust as it produces valid inferences about natural direct and indirect effects in the union model $\mathcal{M}_{\text{union}} = \mathcal{M}_a \cup \mathcal{M}_b \cup \mathcal{M}_c$. Furthermore, as we later show in Section 2.3, the proposed estimators are also locally semiparametric efficient in the sense that they achieve the respective efficiency bounds for estimating the natural direct and indirect effects in $\mathcal{M}_{\text{union}}$, at the intersection submodel $\mathcal{M}_a \cap \mathcal{M}_b \cap \mathcal{M}_c = \mathcal{M}_a \cap \mathcal{M}_c = \mathcal{M}_a \cap \mathcal{M}_b = \mathcal{M}_b \cap \mathcal{M}_c \subset \mathcal{M}_{\text{union}} \subset \mathcal{M}_{\text{nonpar}}$.

Section 3 summarizes a simulation study illustrating the finite sample performance of the various estimators described in Section 2, and Section 4 gives a real data application of these methods. Section 5 describes a strategy to improve the stability of the proposed multiply robust estimator which directly depends on inverse exposure and mediator density weights, when such weights are highly variable, and Section 6 demonstrates the favorable performance of two modified multiply robust estimators in the context of such highly variable weights. In Section 7, we compare the proposed methodology to the prevailing estimators in the literature. Based on this comparison, we conclude that the new approach should generally be preferred because an inference under the proposed method is guaranteed to remain valid under many more data generating laws than an inference based on each of the other existing approaches. In particular, as we argue below the approach of van der Laan and Petersen (2005) is not entirely satisfactory because, despite producing a CAN estimator of the marginal direct effect under the union model $\mathcal{M}_a \cup \mathcal{M}_c$ (and therefore an estimator that is double robust), their estimator requires a correct model for the density of the mediator. Thus, unlike the direct effect estimator developed in this paper, the van der Laan estimator fails to be consistent under the submodel $\mathcal{M}_b \subset \mathcal{M}_{\text{union}}$. Nonetheless, the estimator of van der Laan is in fact locally efficient in model $\mathcal{M}_a \cup \mathcal{M}_c$, provided the model for the mediator's conditional density is either known, or can be efficiently estimated. This property is confirmed in a supplementary online Appendix [Tchetgen Tchetgen and Shpitser (2012)], where we also provide a general map that relates the efficient influence function for model $\mathcal{M}_{\text{union}}$ to the corresponding efficient influence function for model $\mathcal{M}_a \cup \mathcal{M}_c$, assuming an arbitrary parametric or semiparametric model for the mediator conditional density is correctly specified. In Section 8, we describe a novel double robust sensitivity analysis framework to assess the impact on inferences about the natural direct effect, of a departure from the ignorability assumption of the mediator variable. We conclude with a brief discussion.

2. The nonparametric mediation functional.

2.1. *Identification.* Suppose i.i.d. data on $O = (Y, E, M, X)$ is collected for n subjects. Recall that Y is an outcome of interest, E is a binary exposure variable, M is a mediator variable with support $\mathcal{S}$, known to occur subsequently to E and prior to Y and X is a vector of pre-exposure variables with support $\mathcal{X}$ that confound the association between (E, M) and Y. The overarching goal of this paper is to provide some theory of inference about the fundamental functional of mediation analysis which Judea Pearl calls "the mediation causal formula" [Pearl (2011)] and which, expressed on the mean scale, is

$$\begin{aligned}\theta_0 = \iint_{\mathcal{S}\times\mathcal{X}} &\mathbb{E}(Y|E=1, M=m, X=x) \\ &\times f_{M|E,X}(m|E=0, X=x) f_X(x)\, d\mu(m, x),\end{aligned} \tag{2}$$

$f_{M|E,X}$ and f_X are respectively the conditional density of the mediator M given (E, X) and the density of X, and μ is a dominating measure for the distribution of (M, X). Hereafter, to keep with standard statistical parlance, we shall simply refer to θ_0 as the "mediation functional" or "M-functional" since it is formally a functional on the nonparametric statistical model $\mathcal{M}_{\text{nonpar}} = \{F_O(\cdot) : F_O \text{ unrestricted}\}$ of all regular laws F_O of the observed data O that satisfy the positivity assumption given below; that is, $\theta_0 = \theta_0(F_O) : \mathcal{M}_{\text{nonpar}} \to \mathcal{R}$, with $\mathcal{R}$ the real line. The functional θ_0 is of keen interest here because it arises in the estimation of natural direct and indirect effects as we describe next. To do so, we make the consistency assumption.

Consistency:

$$\begin{aligned}&\text{if } E = e, \text{ then } M_e = M \text{ w.p.1} \quad \text{and} \\ &\text{if } E = e \text{ and } M = m, \text{ then } Y_{e,m} = Y \text{ w.p.1}.\end{aligned}$$

In addition, we adopt the sequential ignorability assumption of Imai, Keele and Tingley (2010) which states that for $e, e' \in \{0, 1\}$.

Sequential ignorability:

$$\begin{aligned}\{Y_{e',m}, M_e\} &\perp\!\!\!\perp E|X, \\ Y_{e'm} &\perp\!\!\!\perp M|E=e, X,\end{aligned}$$

where $A \perp\!\!\!\perp B|C$ states that A is independent of B given C; paired with the following:

Positivity:

$$\begin{aligned}f_{M|E,X}(m|E, X) > 0 \qquad &\text{w.p.1 for each } m \in \mathcal{S} \quad \text{and} \\ f_{E|X}(e|X) > 0 \qquad &\text{w.p.1 for each } e \in \{0, 1\}.\end{aligned}$$

Then, under the consistency, sequential ignorability and positivity assumptions, Imai, Keele and Tingley (2010) showed that

$$
\begin{aligned}
\theta_0 &= \mathbb{E}(Y_{1,M_0}) \quad \text{and} \\
\delta_e &\equiv \int_{\mathcal{X}} \mathbb{E}(Y|E=e, X=x) f_X(x)\, d\mu(x) \\
&= \iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=e, M=m, X=x) \\
&\qquad \times f_{M|E,X}(m|E=e, X=x) f_X(x)\, d\mu(m,x) \\
&= \mathbb{E}(Y_e) = \mathbb{E}(Y_{e,M_e}), \qquad e=0,1,
\end{aligned} \tag{3}
$$

so that $\mathbb{E}(Y_{1,M_0})$ and $\mathbb{E}(Y_e)$, $e=0,1$, are identified from the observed data, and so is the mean natural direct effect $\mathbb{E}(Y_{1,M_0}) - \mathbb{E}(Y_0) = \theta_0 - \delta_0$ and the mean natural indirect effect $\mathbb{E}(Y_1) - \mathbb{E}(Y_{1,M_0}) = \delta_1 - \theta_0$. For binary Y, one might alternatively consider the natural direct effect on the risk ratio scale $\mathbb{E}(Y_{1,M_0})/\mathbb{E}(Y_0) = \theta_0/\delta_0$ or on the odds ratio scale $\{\mathbb{E}(Y_{1,M_0})\mathbb{E}(1-Y_0)\}/\{\mathbb{E}(1-Y_{1,M_0})\mathbb{E}(Y_0)\} = \{\theta_0(1-\delta_0)\}/\{\delta_0(1-\theta_0)\}$ and similarly defined natural indirect effects on the risk ratio and odds ratio scales. It is instructive to contrast the expression (2) for $\mathbb{E}(Y_{1,M_0})$ with the expression (3) for $e=1$ corresponding to $\mathbb{E}(Y_1)$, and to note that the two expressions bare a striking resemblance except the density of the mediator in the first expression conditions on the unexposed (with $E=0$), whereas in the second expression, the mediator density is conditional on the exposed (with $E=1$). As we demonstrate below, this subtle difference has remarkable implications for inference.

Pearl (2001) was the first to derive the M-functional $\theta_0 = \mathbb{E}(Y_{1,M_0})$ under a different set of assumptions. Others have since contributed alternative sets of identifying assumptions. In this paper, we have chosen to work under the sequential ignorability assumption of Imai, Keele and Yamamoto (2010), Imai, Keele and Tingley (2010), but note that alternative related assumptions exist in the literature [Robins and Greenland (1992), Pearl (2001), van der Laan and Petersen (2005), Hafeman and Vanderweele (2011)]; however, we note that Robins and Richardson (2012) disagree with the label "sequential ignorability" because its terminology has previously carried a different interpretation in the literature. Nonetheless, the assumption entails two ignorability-like assumptions that are made sequentially. First, given the observed pre-exposure confounders, the exposure assignment is assumed to be ignorable, that is, statistically independent of potential outcomes and potential mediators. The second part of the assumption states that the mediator is ignorable given the observed exposure and pre-exposure confounders. Specifically, the second part of the sequential ignorability assumption is conditional on the observed value of the ignorable treatment and the observed pretreatment confounders. We note that the second part of the sequential ignorability assumption is

particularly strong and must be made with care. This is partly because it is always possible that there might be unobserved variables that confound the relationship between the outcome and the mediator variables, even upon conditioning on the observed exposure and covariates. Furthermore, the confounders X must all be pre-exposure variables; that is, they must precede E. In fact, Avin, Shpitser and Pearl (2005) proved that without additional assumptions, one cannot identify natural direct and indirect effects if there are confounding variables that are affected by the exposure, even if such variables are observed by the investigator [also see Tchetgen Tchetgen and VanderWeele (2012)]. This implies that, similarly to the ignorability of the exposure in observational studies, ignorability of the mediator cannot be established with certainty, even after collecting as many pre-exposure confounders as possible. Furthermore, as Robins and Richardson (2012) point out, whereas the first part of the sequential ignorability assumption could, in principle, be enforced in a randomized study, by randomizing E within levels of X; the second part of the sequential ignorability assumption cannot similarly be enforced experimentally, even by randomization. And thus, for this latter assumption to hold, one must entirely rely on expert knowledge about the mechanism under study. For this reason, it will be crucial in practice to supplement mediation analyses with a sensitivity analysis that accurately quantifies the degree to which results are robust to a potential violation of the sequential ignorability assumption. Later in the paper, we develop a variety of sensitivity analysis techniques that allow the analyst to quantify the degree to which his or her mediation analysis results are robust to a potential violation of the sequential ignorability assumption.

2.2. *Semiparametric efficiency bounds for* $\mathcal{M}_{\text{nonpar}}$. In this section, we derive the efficient influence function for the M-functional θ_0 in $\mathcal{M}_{\text{nonpar}}$. This result is then combined with the efficient influence function for the functional δ_e [Robins, Rotnitzky and Zhao (1994), Hahn (1998)] to obtain the efficient influence function for the natural direct and indirect effects on the mean difference scale. Thus, in the following, we shall use the efficient influence function $S_{\delta e}^{\text{eff,nonpar}}(\delta_e)$ of δ_e which is well known to be

$$\frac{I(E=e)}{f_{E|X}(e|X)}\{Y-\eta(e,e,X)\}+\eta(e,e,X)+\delta_e,$$

where for $e, e^* \in \{0,1\}$, we define

$$\eta(e,e^*,X)=\int_{\mathcal{S}}\mathbb{E}(Y|X,M=m,E=e)\,f_{M|E,X}(m|E=e^*,X)\,d\mu(m),$$

so that $\eta(e,e,X)=\mathbb{E}(Y|X,E=e)$, $e=0,1$.

The following theorem is proved in the Appendix.

THEOREM 1. *Under the consistency, sequential ignorability and positivity assumptions, the efficient influence function of the M-functional* θ_0 *in model* $\mathcal{M}_{\text{nonpar}}$

is given by

$$
\begin{aligned}
&S_{\theta_0}^{\text{eff,nonpar}}(\theta_0)\\
&\quad = s_{\theta_0}^{\text{eff,nonpar}}(O;\theta_0)\\
&\quad = \frac{I\{E=1\}f_{M|E,X}(M|E=0,X)}{f_{E|X}(1|X)f_{M|E,X}(M|E=1,X)}\{Y-\mathbb{E}(Y|X,M,E=1)\}\\
&\qquad + \frac{I(E=0)}{f_{E|X}(0|X)}\{\mathbb{E}(Y|X,M,E=1)-\eta(1,0,X)\}+\eta(1,0,X)-\theta_0,
\end{aligned}
$$

and the efficient influence function of the natural direct and indirect effects on the mean difference scale in model $\mathcal{M}_{\text{nonpar}}$ *are respectively given by*

$$
\begin{aligned}
&S_{\text{NDE}}^{\text{eff,nonpar}}(\theta_0,\delta_0)\\
&\quad = s_{\text{NDE}}^{\text{eff,nonpar}}(O;\theta_0,\delta_0)\\
&\quad = S_{\theta_0}^{\text{eff,nonpar}}(\theta_0)-S_{\delta_0}^{\text{eff,nonpar}}(\delta_0)\\
&\quad = \frac{I\{E=1\}f_{M|E,X}(M|E=0,X)}{f_{E|X}(1|X)f_{M|E,X}(M|E=1,X)}\{Y-\mathbb{E}(Y|X,M,E=1)\}\\
&\qquad + \frac{I(E=0)}{f_{E|X}(0|X)}\{\mathbb{E}(Y|X,M,E=1)-Y-\eta(1,0,X)+\eta(0,0,X)\}\\
&\qquad + \eta(1,0,X)-\eta(0,0,X)-\theta_0+\delta_0,
\end{aligned}
$$

and

$$
\begin{aligned}
&S_{\text{NIE}}^{\text{eff,nonpar}}(\delta_1,\theta_0)\\
&\quad = s_{\text{NIE}}^{\text{eff,nonpar}}(O;\delta_1,\theta_0)\\
&\quad = \frac{I(E=1)}{f_{E|X}(1|X)}\Bigg\{Y-\eta(1,1,X)\\
&\qquad\qquad - \frac{f_{M|E,X}(M|E=0,X)}{f_{M|E,X}(M|E=1,X)}\{Y-\mathbb{E}(Y|X,M,E=1)\}\Bigg\}\\
&\qquad - \frac{I(E=0)}{f_{E|X}(0|X)}\{\mathbb{E}(Y|X,M,E=1)-\eta(1,0,X)\}\\
&\qquad + \eta(1,1,X)-\eta(1,0,X)+\theta_0-\delta_1.
\end{aligned}
$$

Thus, the semiparametric efficiency bound for estimating the natural direct and the natural indirect effects in $\mathcal{M}_{\text{nonpar}}$ *are respectively given by* $\mathbb{E}\{S_{\text{NDE}}^{\text{eff,nonpar}}(\theta_0,\delta_0)^2\}^{-1}$ *and* $\mathbb{E}\{S_{\text{NIE}}^{\text{eff,nonpar}}(\delta_1,\theta_0)^2\}^{-1}$.

Although not presented here, Theorem 1 is easily extended to obtain the efficient influence functions and the respective semiparametric efficiency bounds for the direct and indirect effects on the risk ratio and the odds ratio scales by a straightforward application of the delta method. An important implication of the theorem is that all regular and asymptotically linear (RAL) estimators of θ_0, $\delta_1 - \theta_0$ and $\theta_0 - \delta_0$ in model $\mathcal{M}_{\text{nonpar}}$ share the common influence functions $S_{\theta_0}^{\text{eff,nonpar}}(\theta_0)$, $S_{\text{NDE}}^{\text{eff,nonpar}}(\theta_0, \delta_0)$ and $S_{\text{NIE}}^{\text{eff,nonpar}}(\delta_1, \theta_0)$, respectively. Specifically, any RAL estimator $\widehat{\theta}_0$ of the M-functional θ_0 in model $\mathcal{M}_{\text{nonpar}}$, shares a common asymptotic expansion,

$$n^{1/2}(\widehat{\theta}_0 - \theta_0) = n^{1/2}\mathbb{P}_n S_{\theta_0}^{\text{eff,nonpar}}(\theta_0) + o_P(1),$$

where $\mathbb{P}_n[\cdot] = n^{-1}\sum_i[\cdot]_i$. To illustrate this property of nonparametric RAL estimators, and as a motivation to multiply robust estimation when nonparametric methods are not appropriate, we provide a detailed study of three nonparametric strategies for estimating the M-functional in a simple yet instructive setting in which X and M are both discrete with finite support.

Strategy 1: The first strategy entails obtaining the maximum likelihood estimator upon evaluating the M-functional under the empirical law of the observed data,

$$\widehat{\theta}_0^{\text{ym}} = \mathbb{P}_n \sum_{m \in \mathcal{S}} \widehat{\mathbb{E}}(Y|E=1, M=m, X)\, \widehat{f}_{M|E,X}(m|E=0, X),$$

where $\widehat{f}_{Y|E,M,X}$ and $\widehat{f}_{M|E,X}$ are the empirical probability mass functions, and $\widehat{\mathbb{E}}(Y|E=e, M=m, X=x)$ is the expectation of Y under $\widehat{f}_{Y|E,M,X}$.

Strategy 2: The second strategy is based on the following alternative representation of the M-functional:

$$\begin{aligned}
&\iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=1, M=m, X=x)\, dF_{M|E}(m|E=0, X=x)\, dF_X(x) \\
&\quad = \sum_{e=0}^{1} \iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=1, M=m, X=x) \frac{I(e=0)}{f_{E|X}(e|X=x)}\, dF_{M,E,X}(m, e, x) \\
&\quad = \mathbb{E}\left\{ \frac{I(E=0)}{f_{E|X}(0|X)} \mathbb{E}(Y|E=1, M, X) \right\}.
\end{aligned}$$

Thus, our second estimator takes the form

$$\widehat{\theta}_0^{\text{ye}} = \mathbb{P}_n \left\{ \frac{I(E=0)}{\widehat{f}_{E|X}(0|X)} \widehat{\mathbb{E}}(Y|E=1, M, X) \right\},$$

with $\widehat{f}_{E|X}$ the empirical estimate of the probability mass function $f_{E|X}$.

Strategy 3: The last strategy is based on a third representation of the M-functional

$$\begin{aligned}
&\iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=1, M=m, X=x)\, dF_{M|E}(m|E=0, X=x)\, dF_X(x) \\
&\quad = \sum_{e=0}^{1} \iiint_{\mathcal{Y}\times\mathcal{S}\times\mathcal{X}} y \frac{I(e=1)}{f_{E|X}(e|X=x)} \frac{f_{M|E,X}(M|E=0, X)}{f_{M|E,X}(M|E, X)}\, dF_{Y,M,E,X}(y, m, e, x) \\
&\quad = \mathbb{E}\left\{ Y \frac{I(E=1)}{f_{E|X}(E|X)} \frac{f_{M|E,X}(M|E=0, X)}{f_{M|E,X}(M|E, X)} \right\}.
\end{aligned}$$

Thus, our third estimator takes the form

$$\widehat{\theta}_0^{\text{em}} = \mathbb{P}_n \left\{ Y \frac{I(E=1)}{\widehat{f}_{E|X}(E|X)} \frac{\widehat{f}_{M|E,X}(M|E=0, X)}{\widehat{f}_{M|E,X}(M|E, X)} \right\}.$$

At first glance the three estimators $\widehat{\theta}_0^{\text{em}}$, $\widehat{\theta}_0^{\text{ye}}$ and $\widehat{\theta}_0^{\text{ym}}$ might appear to be distinct; however, we observe that provided the empirical distribution function $\widehat{F}_O = \widehat{F}_{Y|E,M,X} \times \widehat{F}_{M|E,X} \times \widehat{F}_{E|X} \times \widehat{F}_X$ satisfies the positivity assumption, and thus $\widehat{F}_O \in \mathcal{M}_{\text{nonpar}}$, then actually $\widehat{\theta}_0^{\text{em}} = \widehat{\theta}_0^{\text{ye}} = \widehat{\theta}_0^{\text{ym}} = \theta_0(\widehat{F}_O)$ since the three representations agree on the nonparametric model $\mathcal{M}_{\text{nonpar}}$. Therefore we may conclude that these three estimators are in fact asymptotically efficient in $\mathcal{M}_{\text{nonpar}}$ with common influence function $S_{\theta_0}^{\text{eff,nonpar}}(\theta_0)$. Furthermore, from this observation, one further concludes that (asymptotic) inferences obtained using one of the three representations are identical to inferences using either of the other two representations.

At this juncture, we note that the above equivalence no longer applies when as we have previously argued will likely occur in practice, (M, X) contains 3 or more continuous variables and/or X is too high dimensional for models to be saturated or nonparametric, and thus parametric (or semiparametric) models are specified for dimension reduction. Specifically, for such settings, we observe that three distinct modeling strategies are available. Under the first strategy, the estimator $\widehat{\theta}_0^{\text{ym,par}}$ is obtained $\widehat{\theta}_0^{\text{ym}}$ using parametric model estimates $\widehat{\mathbb{E}}^{\text{par}}(Y|E, M, X)$ and $\widehat{f}_{M|E,X}^{\text{par}}(m|E, X)$ instead of their nonparametric counterparts; similarly under the second strategy, the estimator $\widehat{\theta}_0^{\text{ye,par}}$ is obtained similarly to $\widehat{\theta}_0^{\text{ye}}$ using estimates of parametric models $\widehat{\mathbb{E}}^{\text{par}}(Y|E=1, M=m, X)$ and $\widehat{f}_{E|X}^{\text{par}}(e|X)$ and finally, under the third strategy, $\widehat{\theta}_0^{\text{em,par}}$ is obtained similarly to $\widehat{\theta}_0^{\text{em}}$ using $\widehat{f}_{E|X}^{\text{par}}(e|X)$ and $\widehat{f}_{M|E,X}^{\text{par}}(m|E, X)$. Then, it follows that $\widehat{\theta}_0^{\text{ym,par}}$ is CAN under the submodel $\mathcal{M}_a$, but is generally inconsistent if either $\widehat{\mathbb{E}}^{\text{par}}(Y|E, M, X)$ or $\widehat{f}_{M|E,X}^{\text{par}}(m|E, X)$ fails to be consistent. Similarly, $\widehat{\theta}_0^{\text{ye,par}}$ and $\widehat{\theta}_0^{\text{em,par}}$ are, respectively, CAN under the submodels $\mathcal{M}_b$ and $\mathcal{M}_c$, but each estimator generally fails to be consistent outside of the corresponding submodel. In the next section, we propose an approach

that produces a triply robust estimator by combining the above three strategies so that only one of models $\mathcal{M}_a$, $\mathcal{M}_b$ and $\mathcal{M}_c$ needs to be valid for consistency of the estimator.

2.3. *Triply robust estimation*. The proposed triply robust estimator $\widehat{\theta}_0^{\text{triply}}$ solves

$$\mathbb{P}_n \widehat{S}_{\theta_0}^{\text{eff,nonpar}}(\widehat{\theta}_0^{\text{triply}}) = 0,$$

where $\widehat{S}_{\theta_0}^{\text{eff,nonpar}}(\theta)$ is equal to $S_{\theta_0}^{\text{eff,nonpar}}(\theta)$ evaluated at $\{\widehat{\mathbb{E}}^{\text{par}}(Y|E, M, X), \widehat{f}_{M|E,X}^{\text{par}}(m|E, X), \widehat{f}_{E|X}^{\text{par}}(e|X)\}$; that is,

$$\begin{aligned}\widehat{\theta}_0^{\text{triply}} = \mathbb{P}_n\Bigg[&\frac{I\{E=1\}\widehat{f}_{M|E,X}^{\text{par}}(M|E=0,X)}{\widehat{f}_{E|X}^{\text{par}}(1|X)\widehat{f}_{M|E,X}^{\text{par}}(M|E=1,X)} \\ &\times \{Y - \widehat{\mathbb{E}}^{\text{par}}(Y|X, M, E=1)\} \\ &+ \frac{I(E=0)}{\widehat{f}_{E|X}^{\text{par}}(0|X)}\{\widehat{\mathbb{E}}^{\text{par}}(Y|X, M, E=1) \\ &- \widehat{\eta}^{\text{par}}(1,0,X)\} + \widehat{\eta}^{\text{par}}(1,0,X)\Bigg],\end{aligned} \tag{4}$$

is CAN in model $\mathcal{M}_{\text{union}} = \mathcal{M}_a \cup \mathcal{M}_b \cup \mathcal{M}_c$, where

$$\widehat{\eta}^{\text{par}}(e, e^*, X) = \int_{\mathcal{S}} \widehat{\mathbb{E}}^{\text{par}}(Y|X, M=m, E=e)\, \widehat{f}_{M|E,X}^{\text{par}}(m|E=e^*, X)\, d\mu(m).$$

In the next theorem, the estimator in the above display is combined with a doubly robust estimator $\widehat{\delta}_e^{\text{doubly}}$ of δ_e [see van der Laan and Robins (2003) or Tsiatis (2006)], to obtain multiply robust estimators of natural direct and indirect effects, where

$$\widehat{\delta}_e^{\text{doubly}} = \mathbb{P}_n\Bigg[\frac{I(E=e)}{\widehat{f}_{E|X}^{\text{par}}(e|X)}\{Y - \widehat{\eta}^{\text{par}}(e,e,X)\} + \widehat{\eta}^{\text{par}}(e,e,X)\Bigg].$$

To state the result, we set $\widehat{\mathbb{E}}^{\text{par}}(Y|X, M, E) = \mathbb{E}^{\text{par}}(Y|X, M, E; \widehat{\beta}_y) = g^{-1}(\widehat{\beta}_y^T h(X, M, E))$, where g is a known link function, and h is a user specified function of (X, M, E) so that $\mathbb{E}^{\text{par}}(Y|X, M, E; \beta_y) = g^{-1}(\beta_y^T h(X, M, E))$ entails a working regression model for $\mathbb{E}(Y|X, M, E)$, and $\widehat{\beta}_y$ solves the estimating equation

$$0 = \mathbb{P}_n[S_y(\widehat{\beta}_y)] = \mathbb{P}_n\big[h(X, M, E)\big(Y - g^{-1}(\widehat{\beta}_y^T h(X, M, E))\big)\big].$$

Similarly, we set $\widehat{f}_{M|E,X}^{\text{par}}(m|E, X) = f_{M|E,X}^{\text{par}}(m|E, X; \widehat{\beta}_m)$ for $f_{M|E,X}^{\text{par}}(m|E, X; \beta_m)$, a parametric model for the density of $[M|E, X]$ with $\widehat{\beta}_m$, solving

$$0 = \mathbb{P}_n[S_m(\widehat{\beta}_m)] = \mathbb{P}_n\Bigg[\frac{\partial}{\partial \beta_m} \log f_{M|E,X}^{\text{par}}(M|E, X; \widehat{\beta}_m)\Bigg],$$

and we set $\widehat{f}^{\text{par}}_{E|X}(e|X) = f^{\text{par}}_{E|X}(e|X;\widehat{\beta}_e)$ for $f^{\text{par}}_{E|X}(e|X;\beta_e)$, a parametric model for the density of $[E|X]$ with $\widehat{\beta}_e$ solving

$$0 = \mathbb{P}_n[S_e(\widehat{\beta}_e)] = \mathbb{P}_n\left[\frac{\partial}{\partial\beta_e}\log f^{\text{par}}_{E|X}(E|X;\widehat{\beta}_e)\right].$$

THEOREM 2. *Suppose that the assumptions of Theorem 1 hold, and that the regularity conditions stated in the Appendix hold and that β_m, β_e and β_y are variation independent.*

(i) *Mediation functional*: *Then, $\sqrt{n}(\widehat{\theta}_0^{\text{triply}} - \theta_0)$ is RAL under model $\mathcal{M}_{\text{union}}$ with influence function*

$$S^{\text{union}}_{\theta_0}(\theta_0,\beta^*) = S^{\text{eff,nonpar}}_{\theta_0}(\theta_0,\beta^*) - \left.\frac{\partial\mathbb{E}\{S^{\text{eff,nonpar}}_{\theta_0}(\theta_0,\beta)\}}{\partial\beta^T}\right|_{\beta^*}\mathbb{E}\left\{\left.\frac{\partial S_\beta(\beta)}{\partial\beta^T}\right|_{\beta^*}\right\}^{-1}S_\beta(\beta^*),$$

and thus converges in distribution to a $N(0,\Sigma_{\theta_0})$, where

$$\Sigma_{\theta_0}(\theta_0,\beta^*) = \mathbb{E}(S^{\text{union}}_{\theta_0}(\theta_0,\beta^*)^2),$$

with $\beta^T = (\beta_m^T,\beta_e^T,\beta_y^T)$ and $S_\beta(\beta) = (S_m^T(\beta_m), S_e^T(\beta_e), S_y^T(\beta_y))^T$, and with β^* denoting the probability limit of the estimator $\widehat{\beta} = (\widehat{\beta}_m^T,\widehat{\beta}_e^T,\widehat{\beta}_y^T)^T$.

(ii) *Natural direct effect*: *Similarly, $\sqrt{n}(\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}} - (\theta_0 - \delta_0))$ is RAL under model $\mathcal{M}_{\text{union}}$ with influence function $S^{\text{union}}_{\text{NDE}}(\theta_0,\delta_0,\beta^*)$ defined as $S^{\text{union}}_{\theta_0}(\theta_0,\beta^*)$ with $S^{\text{eff,nonpar}}_{\text{NDE}}(\theta_0,\delta_0,\beta^*)$ replacing $S^{\text{eff,nonpar}}_{\theta_0}(\theta_0,\beta^*)$, and asymptotic variance $\Sigma_{\theta_0-\delta_0}(\delta_1,\theta_0,\beta^*)$ defined accordingly.*

(iii) *Natural indirect effect*: *Similarly, $\sqrt{n}(\widehat{\delta}_1^{\text{doubly}} - \widehat{\theta}_0^{\text{triply}} - (\delta_1 - \theta_0))$ is RAL under model M_{union} with influence function $S^{\text{union}}_{\text{NIE}}(\delta_1,\theta_0,\beta^*)$ defined as $S^{\text{union}}_{\theta_0}(\theta_0,\beta^*)$ with $S^{\text{eff,nonpar}}_{\text{NIE}}(\delta_1,\theta_0,\beta^*)$ replacing $S^{\text{eff,nonpar}}_{\theta_0}(\theta_0,\beta^*)$, and asymptotic variance $\Sigma_{\delta_1-\theta_0}(\delta_1,\theta_0,\beta^*)$ defined accordingly.*

(iv) *$\widehat{\theta}_0^{\text{triply}}$, $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\delta}_1^{\text{doubly}} - \widehat{\theta}_0^{\text{triply}}$ are semiparametric locally efficient in the sense that they are RAL under model $\mathcal{M}_{\text{union}}$ and respectively achieve the semiparametric efficiency bound for θ_0, $\theta_0 - \delta_0$, and $\delta_1 - \theta_0$ under model $\mathcal{M}_{\text{union}}$ at the intersection submodel $\mathcal{M}_a \cap \mathcal{M}_b \cap \mathcal{M}_c$, with respective efficient influence functions: $S^{\text{eff,nonpar}}_{\theta_0}(\theta_0,\beta^*)$, $S^{\text{eff,nonpar}}_{\text{NDE}}(\theta_0,\delta_0,\beta^*)$ and $S^{\text{eff,nonpar}}_{\text{NIE}}(\delta_1,\theta_0,\beta^*)$.*

Empirical versions of $\Sigma_{\theta_0-\delta_0}(\delta_1,\theta_0,\beta^*)$ and $\Sigma_{\delta_1-\theta_0}(\delta_1,\theta_0,\beta^*)$ are easily obtained, and the corresponding Wald-type confidence intervals can be used to make formal inferences about natural direct and indirect effects. It is also straightforward to extend the approach to the risk ratio and odds ratio scales for binary Y. By a theorem due to Robins and Rotnitzky (2001), part (iv) of the theorem implies that when all models are correct, $\widehat{\theta}_0^{\text{triply}}$, $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$ and

$\widehat{\delta}_1^{\text{doubly}} - \widehat{\theta}_0^{\text{triply}}$ are semiparametric efficient in model $\mathcal{M}_{\text{nonpar}}$ at the intersection submodel $\mathcal{M}_a \cap \mathcal{M}_b \cap \mathcal{M}_c$.

3. A simulation study of estimators of direct effect. In this section, we report a simulation study which illustrates the finite sample performance of the various estimators described in previous sections. We generated 1000 samples of size $n = 600, 1000$ from the following model:

(Model.X) $X_1 \sim \textit{Bernoulli}(0.4)$; $[X_2|X_1] \sim \textit{Bernoulli}(0.3 + 0.4X_1)$;
$[X_3|X_1, X_2] \sim -0.024 - 0.4X_1 + 0.4X_2 + N(0, 1)$;

(Model.E) $[E|X_1, X_2, X_3] \sim \textit{Bernoulli}([1 + \exp\{-(0.4 + X_1 - X_2 + 0.1X_3 - 1.5X_1X_3)\}]^{-1})$;

(Model.M) $[M|E, X_1, X_2, X_3] \sim \textit{Bernoulli}([1 + \exp\{-(0.5 - X_1 + 0.5X_2 - 0.9X_3 + E - 1.5X_1X_3)\}]^{-1})$;

(Model.Y) $[Y|M, E, X_1, X_2, X_3] \sim 1 + 0.2X_1 + 0.3X_2 + 1.4X_3 - 2.5E - 3.5M + 5EM + N(0, 1)$.

We then evaluated the performance of the following four estimators of the natural direct effect $\widehat{\theta}_0^{\text{em}} - \widehat{\delta}_0^{\text{doubly}}$, $\widehat{\theta}_0^{\text{ye}} - \widehat{\delta}_0^{\text{doubly}}$, $\widehat{\theta}_0^{\text{ym}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$. Note that the doubly robust estimator $\widehat{\delta}_0^{\text{doubly}}$ was used throughout to estimate $\delta_0 = \mathbb{E}(Y_0)$. To assess the impact of modeling error, we evaluated these estimators in four separate scenarios. In the first scenario, all models were correctly specified, whereas the remaining three scenarios respectively mis-specified only one of Model E, Model M and Model Y. In order to mis-specify Model E and Model M, we respectively left out the X_1X_3 interaction when fitting each model, and we assumed an incorrect log–log link function. The incorrect model for Y simply assumed no EM interaction.

Tables 1 and 2 summarize the simulation results which largely agree with the theory developed in the previous sections. Mainly, all proposed estimators performed well at both moderate and large sample sizes in the absence of modeling error. Furthermore, under the partially mis-specified model in which Model.Y was incorrect, both estimators, $\widehat{\theta}_0^{\text{ye}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\theta}_0^{\text{ym}} - \widehat{\delta}_0^{\text{doubly}}$, showed significant bias irrespective of sample size, while $\widehat{\theta}_0^{\text{em}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$ both performed well. Similarly when Model M was incorrect, the estimators $\widehat{\theta}_0^{\text{em}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\theta}_0^{\text{ym}} - \widehat{\delta}_0^{\text{doubly}}$ resulted in large bias, when compared to the relatively small bias of $\widehat{\theta}_0^{\text{ye}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$. Finally, mis-specifying Model E lead to estimators $\widehat{\theta}_0^{\text{ye}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\theta}_0^{\text{em}} - \widehat{\delta}_0^{\text{doubly}}$ that were significantly more biased than the estimators $\widehat{\theta}_0^{\text{ym}} - \widehat{\delta}_0^{\text{doubly}}$ and $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$. Interestingly, the efficiency loss of the multiply robust estimator remained relatively small when compared to the consistent nonrobust estimator under the various scenarios, suggesting that, at least in this simulation study, the benefits of robustness appear to outweigh the loss of efficiency.

TABLE 1
Simulation results $n = 600$

		$\mathcal{M}_{\mathbf{ym}}$	$\mathcal{M}_{\mathbf{ye}}$	$\mathcal{M}_{\mathbf{em}}$	$\mathcal{M}_{\mathbf{union}}$
All correct	bias	0.002	0.008	0.002	0.005
	MC s.e.*	0.005	0.007	0.006	0.006
Y wrong	bias	−0.500	−0.500	0.0001	0.004
	MC s.e.	0.005	0.006	0.006	0.006
M wrong	bias	0.038	0.008	−0.054	0.003
	MC s.e.	0.005	0.007	0.006	0.006
E wrong	bias	0.003	0.027	0.059	0.004
	MC s.e.	0.005	0.005	0.005	0.005

$\mathcal{M}_{\text{ym}}$: $\widehat{\theta}_0^{\text{ym}} - \widehat{\delta}_0^{\text{doubly}}$; $\mathcal{M}_{\text{ye}}$: $\widehat{\theta}_0^{\text{ye}} - \widehat{\delta}_0^{\text{doubly}}$; $\mathcal{M}_{\text{em}}$: $\widehat{\theta}_0^{\text{em}} - \widehat{\delta}_0^{\text{doubly}}$; $\mathcal{M}_{\text{union}}$: $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$.
*Monte Carlo standard error.

4. A data application. In this section, we illustrate the methods in a real world application from the psychology literature on mediation. We re-analyze data from The Job Search Intervention Study (JOBS II) also analyzed by Imai, Keele and Tingley (2010). JOBS II is a randomized field experiment that investigates the efficacy of a job training intervention on unemployed workers. The program is designed not only to increase reemployment among the unemployed but also to enhance the mental health of the job seekers. In the study, 1801 unemployed workers received a pre-screening questionnaire and were then randomly assigned to treatment and control groups. The treatment group with $E = 1$ participated in job skills workshops in which participants learned job search skills and coping strategies for dealing with setbacks in the job search process. The control group with $E = 0$ received a booklet describing job search tips. An analysis considers

TABLE 2
Simulation results $n = 1000$

		$\mathcal{M}_{\mathbf{ym}}$	$\mathcal{M}_{\mathbf{ye}}$	$\mathcal{M}_{\mathbf{em}}$	$\mathcal{M}_{\mathbf{union}}$
All correct	bias	0.001	0.009	0.001	0.001
	MC s.e.*	0.004	0.005	0.004	0.004
Y wrong	bias	−0.484	−0.484	0.003	0.003
	MC s.e.	0.004	0.004	0.004	0.004
M wrong	bias	0.136	−0.008	0.056	0.01
	MC s.e.	0.004	0.05	0.004	0.01
E wrong	bias	0.001	−0.024	−0.054	0.001
	MC s.e.	0.004	0.004	0.004	0.004

$\mathcal{M}_{\text{ym}}$: $\widehat{\theta}_0^{\text{ym}} - \widehat{\delta}_0^{\text{doubly}}$; $\mathcal{M}_{\text{ye}}$: $\widehat{\theta}_0^{\text{ye}} - \widehat{\delta}_0^{\text{doubly}}$; $\mathcal{M}_{\text{em}}$: $\widehat{\theta}_0^{\text{em}} - \widehat{\delta}_0^{\text{doubly}}$; $\mathcal{M}_{\text{union}}$: $\widehat{\theta}_0^{\text{triply}} - \widehat{\delta}_0^{\text{doubly}}$.
*Monte Carlo standard error.

TABLE 3
Estimated causal effects of interest using the job search intervention study data

		$\mathcal{M}_{\mathbf{ym}}$	$\mathcal{M}_{\mathbf{ye}}$	$\mathcal{M}_{\mathbf{em}}$	$\mathcal{M}_{\mathbf{union}}$
Direct effect	Estimate	−0.0310	−0.0310	0.0280	−0.0409
	s.e.*	0.0124	0.0620	0.0465	0.0217
Indirect effect	Estimate	−0.0160	−0.0160	−0.0750	−0.0070
	s.e.*	0.0372	0.0620	0.0434	0.0217

*Nonparametric bootstrap standard errors.

a continuous outcome measure Y of depressive symptoms based on the Hopkins Symptom Checklist [Imai, Keele and Tingley (2010)]. In the JOBS II data, a continuous measure of job search self-efficacy represented the hypothesized mediating variable M. The data also included baseline covariates X measured before administering the treatment including: pretreatment level of depression, education, income, race, marital status, age, sex, previous occupation, and the level of economic hardship.

Note that by randomization, the density of $[E|X]$ was known by design not to depend on covariates, and therefore its estimation is not prone to modeling error. The continuous outcome and mediator variables were modeled using linear regression models with Gaussian error, with main effects for (E, M, X) included in the outcome regression and main effects for (E, X) included in the mediator regression. Table 3 summarizes results obtained using $\widehat{\theta}_0^{\text{em}}$, $\widehat{\theta}_0^{\text{ye}}$, $\widehat{\theta}_0^{\text{ym}}$ and $\widehat{\theta}_0^{\text{triply}}$ together with $\widehat{\delta}_e^{\text{doubly}}$, $e = 0, 1$, to estimate the direct and indirect effects of the treatment.

Point estimates of both natural direct and indirect effects closely agreed under models $\mathcal{M}_{\text{ym}}$ and $\mathcal{M}_{\text{ye}}$, and also agreed with the results of Imai, Keele and Tingley (2010). We should note that inferences under our choice of $\mathcal{M}_{\text{ym}}$ are actually robust to the normality assumption and, as in Imai, Keele and Tingley (2010), only require that the mean structure of $[Y|E, M, X]$ and $[M|E, X]$ is correct. In contrast, inferences under model $\mathcal{M}_{\text{em}}$ require a correct model for the mediator density. This distinction may partly explain the apparent disagreement in the estimated direct effect under $\mathcal{M}_{\text{em}}$ when compared to the other methods, also suggesting that the Gaussian error model for M is not entirely appropriate. The multiply robust estimate of the natural direct effect is consistent with estimates obtained under models $\mathcal{M}_{\text{ym}}$ and $\mathcal{M}_{\text{ye}}$, and is statistically significant, suggesting that the intervention may have beneficial direct effects on participants' mental health; while the multiply robust approach suggests a much smaller indirect effect than all other estimators although none achieved statistical significance.

5. Improving the stability of $\widehat{\boldsymbol{\theta}}_{\mathbf{0}}^{\mathbf{triply}}$ when weights are highly variable. The triply robust estimator $\widehat{\theta}_0^{\text{triply}}$ which involves inverse probability weights for the exposure and mediator variables, clearly relies on the positivity assumption, for good

finite sample performance. But as recently shown by Kang and Schafer (2007) in the context of missing outcome data, a practical violation of positivity in data analysis can severely compromise inferences based on such methodology; although their analysis did not directly concern the M-functional θ_0. Thus, it is crucial to critically examine, as we do below in a simulation study, the extent to which the various estimators discussed in this paper are susceptible to a practical violation of the positivity assumption, and to consider possible approaches to improve the finite sample performance of these estimators in the context of highly variable empirical weights. Methodology to enhance the finite sample behavior of $\widehat{\delta}_j^{\text{doubly}}$ is well studied in the literature and is not considered here; see, for example, Robins et al. (2007), Cao, Tsiatis and Davidian (2009) and Tan (2010). We first describe an approach to enhance the finite sample performance of $\widehat{\theta}_0^{\text{triply}}$, particularly in the presence of highly variable empirical weights. To focus the exposition, we only consider the case of a continuous Y and a binary M, but in principle, the approach could be generalized to a more general setting. The proposed enhancement involves two modifications.

The first modification adapts to the mediation context, an approach developed for the missing data context (and for the estimation of total effects) in Robins et al. (2007). The basic guiding principle of the approach is to carefully modify the estimation of the outcome and mediator models in order to ensure that the triply robust estimator given by equation (4) has the simple M-functional representation

$$\widehat{\theta}_0^{\text{triply},\dagger} = \mathbb{P}_n\{\widehat{\eta}^{\text{par},\dagger}(1, 0, X)\},$$

where $\widehat{\eta}^{\text{par},\dagger}(1, 0, X)$ is carefully estimated to ensure multiple robustness. The reason for favoring an estimator with the above representation is that it is expected to be more robust to practical positivity violation because it does not directly depend on inverse probability weights. However, as we show next, to ensure multiple robustness, estimation of η^{par} involves inverse probability weights, and therefore, $\widehat{\theta}_0^{\text{triply},\dagger}$ indirectly depends on such weights. Our strategy involves a second step to minimize the potential impact of this indirect dependence on weights.

In the following, we assume, to simplify the exposition, that a simple linear model is used:

$$\mathbb{E}^{\text{par}}(Y|X, M, E = 1) = \mathbb{E}^{\text{par}}(Y|X, M, 1; \beta_y) = [1, X^T, M]\beta_y.$$

Then, similar to Robins et al. (2007), one can verify that the above M-functional representation of a triply robust estimator is obtained by estimating $f^{\text{par}}_{M|E,X}(M|E = 0, X)$ with $\widehat{f}^{\text{par},\dagger}_{M|E,X}(M|E = 0, X)$ obtained via weighted logistic regression in the unexposed-only, with weight $\widehat{f}^{\text{par}}_{E|X}(0|X)^{-1}$; and by estimating $\mathbb{E}^{\text{par}}(Y|X, M, E = 1)$ using weighted OLS of Y on (M, X) in the exposed-only, with weight

$$\widehat{f}^{\text{par},\dagger}_{M|E,X}(M|E = 0, X)\{\widehat{f}^{\text{par}}_{E|X}(1|X)\widehat{f}^{\text{par},\dagger}_{M|E,X}(M|E = 1, X)\}^{-1};$$

provided that both working models include an intercept. The second enhancement to minimize undue influence of variable weights on the M-functional estimator, entails using $\widehat{f}^{\text{par},\dagger}_{E|X}$ in the previous step instead of $\widehat{f}^{\text{par}}_{E|X}$, where

$$\text{logit}\, \widehat{f}^{\text{par},\dagger}_{E|X}(1|X) = \text{logit}\, \widehat{f}^{\text{par}}_{E|X}(1|X) + \widehat{C}_1$$

with

$$\widehat{C}_1 = -\log\big(1 - \mathbb{P}_n(E)\big) + \log\big(\mathbb{P}_n[E\, \widehat{f}^{\text{par}}_{E|X}(0|X)/\widehat{f}^{\text{par}}_{E|X}(1|X)]\big).$$

This second modification ensures a certain boundedness property of inverse propensity score-weighting. Specifically, for any bounded function $R = r(Y, M)$ of Y and M; consider for a moment the goal of estimating the counterfactual mean $\mathbb{E}\{r(Y_1, M_1)\}$; then it is well known that even though R is bounded, the simple inverse-probability weighting estimator $\mathbb{P}_n\{ER\widehat{f}^{\text{par}}_{E|X}(1|X)^{-1}\}$ could easily be unbounded, particularly if positivity is practically violated. In contrast, as we show next, the estimator $\mathbb{P}_n\{ER\widehat{f}^{\text{par},\dagger}_{E|X}(1|X)^{-1}\}$ is generally bounded. To see why, note that

$$\begin{aligned}\mathbb{P}_n\{ER\widehat{f}^{\text{par},\dagger}_{E|X}(1|X)^{-1}\} &= \mathbb{P}_n\{ER\widehat{f}^{\text{par},\dagger}_{E|X}(0|X)\widehat{f}^{\text{par},\dagger}_{E|X}(1|X)^{-1}\} + \mathbb{P}_n\{R\} \\ &= \mathbb{P}_n\left\{R\frac{E\,\widehat{f}^{\text{par}}_{E|X}(0|X)\widehat{f}^{\text{par}}_{E|X}(1|X)^{-1}}{\mathbb{P}_n[E\,\widehat{f}^{\text{par}}_{E|X}(0|X)\widehat{f}^{\text{par}}_{E|X}(1|X)^{-1}]}\big(1 - \mathbb{P}_n(E)\big)\right\} \\ &\quad + \mathbb{P}_n\{R\}\end{aligned}$$

which is bounded since the second term is bounded, and the first term is a convex combination of bounded variables, and therefore is also bounded. Furthermore, $\mathbb{P}_n[E\,\widehat{f}^{\text{par},\dagger}_{E|X}(0|X)\widehat{f}^{\text{par},\dagger}_{E|X}(1|X)^{-1}]$ converges in probability to $(1 - \mathbb{E}(E))$ provided that $\widehat{f}^{\text{par}}_{E|X}$ converges to $f_{E|X}$, ensuring that the expression in the above display is consistent for $\mathbb{E}\{r(Y_1, M_1)\}$. The nonparametric bootstrap is most convenient for inference using $\widehat{f}^{\text{par},\dagger}_{E|X}$.

In the next section, we study, in the context of highly variable weights, the behavior of our previous estimators of θ_0, together with that of the enhanced estimators $\widehat{\theta}^{\text{triply},\dagger,j}_0 = \mathbb{P}_n\{\widehat{\eta}^{\text{par},\dagger,j}(1, 0, X)\}$, $j = 1, 2$, where $\widehat{\eta}^{\text{par},\dagger,1}$ is constructed as described above using $\widehat{f}^{\text{par}}_{E|X}$, and $\widehat{\eta}^{\text{par},\dagger,2}$ uses $\widehat{f}^{\text{par},\dagger}_{E|X}$.

6. A simulation study where positivity is practically violated. We adapted to the mediation setting, the missing data simulation scenarios in Kang and Schafer (2007) which were specifically designed so that, when misspecified, working models are nonetheless nearly correct, but yield highly variable inverse probability weights with practical positivity violation in the context of estimation. We generated 1000 samples of size $n = 200$, 1000 from the following model:

$$\text{(Model.X)}\quad Z = Z_1, Z_2, Z_3, Z_4 \overset{\text{i.i.d.}}{\sim} N(0, 1);\ X_1 = \exp(Z_1/2);$$

$X_2 = Z_2/\{1+\exp(Z_1)\} + 10$; $X_3 = (Z_1 Z_3/25 + 0.6)^3$ and $X_4 = (Z_2 + Z_4 + 20)^2$, so that Z may be expressed in terms of X.

(Model.E) $[E|X_1, X_2, X_3] \sim Bernoulli([1+\exp\{(Z_1 - 0.5Z_2 + 0.25Z_3 + 0.1Z_4)\}]^{-1})$;

(Model.M) $[M|E, X_1, X_2, X_3] \sim Bernoulli([1+\exp\{-(0.5 - Z_1 + 0.5Z_2 - 0.9Z_3 + Z_4 - 1.5E)\}]^{-1})$;

(Model.Y) $[Y|M, E, X_1, X_2, X_3] \sim 210 + 27.4Z_1 + 13.7Z_3 + 13.7Z_3 + M + E + N(0,1)$.

Correctly specified working models were thus achieved when an additive linear regression of Y on Z, a logistic regression of M with linear predictor additive in Z and E and a logistic regression of E with linear predictor additive in the Z, respectively. Incorrect specification involved fitting these models with X replacing Z, which produces higly variable weights. For instance, an estimated propensity score as small as 5.5×10^{-33} occurred in the simulation study reflecting an effective violation of positivity; similarly, a mediator predicted probability as small as 3×10^{-20} also occured in the simulation study.

Tables 4 and 5 summarize simulation results for $\widehat{\theta}_0^{\text{ym}}$, $\widehat{\theta}_0^{\text{ye}}$, $\widehat{\theta}_0^{\text{em}}$, $\widehat{\theta}_0^{\text{triply}}$, $\widehat{\theta}_0^{\text{triply},\dagger,1}$ and $\widehat{\theta}_0^{\text{triply},\dagger,2}$. When all three working models are correct, all estimators perform well in terms of bias, but there are clear differences between the estimators in

TABLE 4
Simulation results $n = 200$

		$\mathcal{M}_{\text{ym}}$	$\mathcal{M}_{\text{ye}}$	$\mathcal{M}_{\text{em}}$	$\mathcal{M}_{\text{union}}$	$\mathcal{M}_{\text{union}}^{\dagger,1}$	$\mathcal{M}_{\text{union}}^{\dagger,2}$
All correct	bias	0.001	−0.207	0.498	0.003	−0.08	−0.079
	MC s.e.*	2.614	8.333	20.214	2.6151	2.6155	2.6153
Y wrong	bias	−9.87	−10.221	0.498	−0.147	−0.502	−0.202
	MC s.e.	3.322	10.539	20.214	4.461	3.177	3.141
M wrong	bias	−0.033	−0.207	−9.497	0.001	0.046	0.046
	MC s.e.	2.613	8.333	15.376	2.615	2.614	2.614
E wrong	bias	−0.001	0.132	210.450	0.066	−0.089	−0.087
	MC s.e.	2.614	4.373	2336.92	4.891	2.619	2.615
Y, E wrong	bias	−9.869	−13.535	210.454	−33.090	−1.4609	−2.487
	MC s.e.	3.322	5.256	2336.92	375.334	5.187	4.245
Y, M wrong	bias	−9.355	−10.220	−9.496	−4.346	−3.579	−3.579
	MC s.e.	3.224	10.539	15.376	3.912	3.480	3.441
E, M wrong	bias	−0.032	0.132	205.060	0.088	−0.001	-3.77×10^{-5}
	MC s.e.	2.614	4.373	2289.788	4.763	2.623	2.618
Y, E, M wrong	bias	−9.355	−13.535	205.060	−37.757	−4.223	−5.253
	MC s.e.	3.224	5.356	2289.78	379.122	5.835	4.828

$\mathcal{M}_{\text{ym}}$: $\widehat{\theta}_0^{\text{ym}}$; $\mathcal{M}_{\text{ye}}$: $\widehat{\theta}_0^{\text{ye}}$; $\mathcal{M}_{\text{em}}$: $\widehat{\theta}_0^{\text{em}}$; $\mathcal{M}_{\text{union}}$: $\widehat{\theta}_0^{\text{triply}}$; $\mathcal{M}_{\text{union}}^{\dagger,1}$: $\widehat{\theta}_0^{\text{triply},\dagger,1}$; $\mathcal{M}_{\text{union}}^{\dagger,2}$: $\widehat{\theta}_0^{\text{triply},\dagger,2}$.
*Monte Carlo standard error.

TABLE 5
Simulation results $n = 1000$

		$\mathcal{M}_{\mathbf{ym}}$	$\mathcal{M}_{\mathbf{ye}}$	$\mathcal{M}_{\mathbf{em}}$	$\mathcal{M}_{\mathbf{union}}$	$\mathcal{M}^{\dagger,\mathbf{1}}_{\mathbf{union}}$	$\mathcal{M}^{\dagger,\mathbf{2}}_{\mathbf{union}}$
All correct	bias	0.0324	0.004	−0.106	0.034	−0.047	−0.047
	MC s.e.*	1.136	3.06	6.490	1.136	1.137	1.137
Y wrong	bias	−10.256	−10.305	−0.106	0.063	−0.147	−0.148
	MC s.e.	1.675	4.005	6.490	1.769	1.419	1.407
M wrong	bias	-5×10^{-4}	0.004	−9.706	0.033	0.076	0.076
	MC s.e.	1.136	3.060	5.395	1.137	1.137	1.135
E wrong	bias	0.032	0.135	2.4×10^6	1908.76	−0.038	−0.030
	MC s.e.	1.136	1.794	4.3×10^7	53911.63	1.400	1.242
Y, *E* wrong	bias	−10.256	−14.011	2.4×10^6	-1.1×10^6	6.201	1.024
	MC s.e.	1.675	2.386	4.3×10^7	2.1×10^7	9.406	5.097
Y, *M* wrong	bias	−9.705	−10.305	−9.706	−4.216	−3.555	−3.557
	MC s.e.	1.626	4.004	5.395	1.667	1.527	1.510
E, *M* wrong	bias	5.7×10^{-4}	0.135	2.5×10^6	2034.83	0.0539	0.0599
	MC s.e.	1.136	1.794	4.6×10^7	56090.10	1.429	1.272
Y, *E*, *M* wrong	bias	−9.075	−14.011	2.5×10^6	-1.2×10^6	4.659	−0.755
	MC s.e.	1.626	2.386	4.6×10^7	2.2×10^7	10.121	5.910

$\mathcal{M}_{\text{ym}}$: $\widehat{\theta}_0^{\text{ym}}$; $\mathcal{M}_{\text{ye}}$: $\widehat{\theta}_0^{\text{ye}}$; $\mathcal{M}_{\text{em}}$: $\widehat{\theta}_0^{\text{em}}$; $\mathcal{M}_{\text{union}}$: $\widehat{\theta}_0^{\text{triply}}$; $\mathcal{M}^{\dagger,1}_{\text{union}}$: $\widehat{\theta}_0^{\text{triply},\dagger,1}$; $\mathcal{M}^{\dagger,2}_{\text{union}}$: $\widehat{\theta}_0^{\text{triply},\dagger,2}$.
*Monte Carlo standard error.

terms of efficiency. In fact, $\widehat{\theta}_0^{\text{ym}}$, $\widehat{\theta}_0^{\text{triply}}$, $\widehat{\theta}_0^{\text{triply},\dagger,1}$ and $\widehat{\theta}_0^{\text{triply},\dagger,2}$ have comparable efficiency for $n = 200, 1000$, but $\widehat{\theta}_0^{\text{ye}}$, $\widehat{\theta}_0^{\text{em}}$ is far more variable. Moreover, under mis-specification of a single model, $\widehat{\theta}_0^{\text{triply}}$, $\widehat{\theta}_0^{\text{triply},\dagger,1}$ and $\widehat{\theta}_0^{\text{triply},\dagger,2}$ remain nearly unbiased, and for the most part substantially more efficient than the corresponding consistent estimator in $\{\widehat{\theta}_0^{\text{ym}}, \widehat{\theta}_0^{\text{ye}}, \widehat{\theta}_0^{\text{em}}\}$. When at least two models are mis-specified, the multiply robust estimators $\widehat{\theta}_0^{\text{triply}}$, $\widehat{\theta}_0^{\text{triply},\dagger,1}$ and $\widehat{\theta}_0^{\text{triply},\dagger,2}$ generally outperform the other estimators, although $\widehat{\theta}_0^{\text{triply}}$ occasionally succumbs to the unstable weights resulting in disastrous mean squared error; see Table 5 when Model M and Model E are both incorrect. In contrast, $\widehat{\theta}_0^{\text{triply},\dagger,2}$ generally improves on $\widehat{\theta}_0^{\text{triply},\dagger,1}$ which generally outperforms $\widehat{\theta}_0^{\text{triply}}$ and for the most part $\widehat{\theta}_0^{\text{triply},\dagger,1}$ and $\widehat{\theta}_0^{\text{triply},\dagger,2}$ appear to eliminate any possible deleterious impact of highly variable weights.

7. A comparison to some existing estimators. In this section, we briefly compare the proposed approach to some existing estimators in the literature. Perhaps the most common approach for estimating direct and indirect effects when *Y* is continuous uses a system of linear structural equations; whereby, a linear structural equation for the outcome, given the exposure, the mediator and the confounders, is combined with a linear structural equation for the mediator, given

the exposure and confounders, to produce an estimator of natural direct and indirect effects. The classical approach of Baron and Kenny (1986) is a particular instance of this approach. In recent work, mainly motivated by Pearl's mediation functional, several authors [Imai, Keele and Tingley (2010), Imai, Keele and Yamamoto (2010), Pearl (2011), VanderWeele (2009), Vanderweele and Vansteelandt (2010)] have demonstrated how the simple linear structural equation approach generalizes to accommodate both, the presence of an interaction between exposure and mediator variables, and a nonlinear link function, either in the regression model for the outcome, or in the regression model for the mediator, or both. In fact, when the effect of confounders is also modeled in such structural equations, inferences based on the latter can be viewed as special instances of inferences obtained under a particular specification of model $\mathcal{M}_a$ for the outcome and the mediator densities. And thus, as previously shown in the simulations, an estimator obtained under a system of structural equations will generally fail to produce a consistent estimator of natural direct and indirect effects when model $\mathcal{M}_a$ is incorrect, whereas, by using the proposed multiply robust estimator, valid inferences can be recovered under the union model $\mathcal{M}_b \cup \mathcal{M}_c$, even if $\mathcal{M}_a$ fails.

A notable improvement on the system of structural equations approach is the double robust estimator of a natural direct effect due to van der Laan and Petersen (2005). Their estimator solves the estimating equation constructed using an empirical version of $S^{\text{eff},\mathcal{M}_a\cup\mathcal{M}_c}_{\text{NDE,singleton}}(\theta_0, \delta_0)$ given in the online Appendix. They show their estimator remains CAN in the larger submodel $\mathcal{M}_a \cup \mathcal{M}_c$ and therefore, they can recover valid inferences even when the outcome model is incorrect, provided both the exposure and mediator models are correct. Unfortunately, the van der Laan estimator is still not entirely satisfactory because unlike the proposed multiply robust estimator, it requires that the model for the mediator density is correct. Nonetheless, if the mediator model is correct, the authors establish that their estimator achieves the efficiency bound for model $\mathcal{M}_a \cup \mathcal{M}_c$ at the intersection submodel $\mathcal{M}_a \cap \mathcal{M}_c$ where all models are correct; and thus it is locally semiparametric efficient in $\mathcal{M}_a \cup \mathcal{M}_c$. Interestingly, as we report in the online supplement, the semiparametric efficiency bounds for models $\mathcal{M}_a \cup \mathcal{M}_c$ and $\mathcal{M}_a \cup \mathcal{M}_b \cup \mathcal{M}_c$ are distinct, because the density of the mediator variable is not ancillary for inferences about the M-functional. Thus, any restriction placed on the mediator's conditional density can, when correct, produce improvements in efficiency. This is in stark contrast with the role played by the density of the exposure variable, which as in the estimation of the marginal causal effect, remains ancillary for inferences about the M-functional and thus the efficiency bound for the latter is unaltered by any additional information on the former [Robins, Rotnitzky and Zhao (1994)]. In the online Appendix, we provide a general functional map that relates the efficient influence function for the larger model $\mathcal{M}_a \cup \mathcal{M}_b \cup \mathcal{M}_c$ to the efficient influence for the smaller model $\mathcal{M}_a \cup \mathcal{M}_c$ where the model for the mediator is either parametric or semiparametric. Our map is instructive because it makes explicit using

simple geometric arguments, the information that is gained from increasing restrictions on the law of the mediator. In the online Appendix, we illustrate the map by recovering the efficient influence function of van der Laan and Petersen in the case of a singleton model (i.e., a known conditional density) for the mediator and in the case of a parametric model for the mediator.

8. A semiparametric sensitivity analysis. We describe a semiparametric sensitivity analysis framework to assess the extent to which a violation of the ignorability assumption for the mediator might alter inferences about natural direct and indirect effects. Although only results for the natural direct effect are given here, the extension for the indirect effect is easily deduced from the presentation. Let

$$t(e,m,x) = \mathbb{E}[Y_{1,m}|E=e, M=m, X=x] - \mathbb{E}[Y_{1,m}|E=e, M\neq m, X=x],$$

then

$$Y_{e',m} \not\perp\!\!\!\perp M|E=e, X,$$

that is, a violation of the ignorability assumption for the mediator variable, generally implies that

$$t(e,m,x) \neq 0 \qquad \text{for some } (e,m,x).$$

Thus, we proceed as in Robins, Rotnitzky and Scharfstein (2000), and propose to recover inferences by assuming the selection bias function $t(e,m,x)$ is known, which encodes the magnitude and direction of the unmeasured confounding for the mediator. In the following, the support of M, $\mathcal{S}$ is assumed to be finite. To motivate the proposed approach, suppose for the moment that $f_{M|E,X}(M|E,X)$ is known; then under the assumption that the exposure is ignorable given X, we show in the Appendix that

$$\begin{aligned}
&\mathbb{E}[Y_{1,m}|M_0=m, X=x]\\
&\quad= \mathbb{E}[Y_{1,m}|E=0, M=m, X=x]\\
&\quad= \mathbb{E}[Y|E=1, M=m, X=x] - t(1,m,x)\big(1 - f_{M|E,X}(m|E=1, X=x)\big)\\
&\qquad + t(0,m,x)\big(1 - f_{M|E,X}(m|E=0, X=x)\big),
\end{aligned}$$

and therefore the M-functional is identified by

$$\begin{aligned}
&\sum_{m\in\mathcal{S}} \mathbb{E}\big\{\mathbb{E}[Y|E=1, M=m, X] - t(1,m,X)\big(1 - f_{M|E,X}(m|E=1,X)\big)\\
&\qquad + t(0,m,X)\big(1 - f_{M|E,X}(m|E=0,X)\big)\big\}\\
&\qquad \times f_{M|E,X}(m|E=0,X),
\end{aligned} \tag{5}$$

which is equivalently represented as

$$\mathbb{E}\Bigg[\frac{I\{E=1\}f_{M|E,X}(M|E=0,X)}{f_{E|X}(1|X)f_{M|E,X}(M|E=1,X)} \times \big\{Y - t(1,M,X)\big(1 - f_{M|E,X}(m|E=1,X)\big) + t(0,M,X)\big(1 - f_{M|E,X}(M|E=0,X)\big)\big\}\Bigg]. \tag{6}$$

Below, these two equivalent representations, (5) and (6), are carefully combined to obtain a double robust estimator of the M-functional, assuming $t(\cdot,\cdot,\cdot)$ is known. A sensitivity analysis is then obtained by repeating this process and reporting inferences for each choice of $t(\cdot,\cdot,\cdot)$ in a finite set of user-specified functions $\mathcal{T} = \{\, t_\lambda(\cdot,\cdot,\cdot) : \lambda\}$ indexed by a finite dimensional parameter λ with $t_0(\cdot,\cdot,\cdot) \in \mathcal{T}$ corresponding to the unmeasured confounding assumption, that is, $t_0(\cdot,\cdot,\cdot) \equiv 0$. Throughout, the model $f^{\text{par}}_{M|E,X}(\cdot|E,X;\beta_m)$ for the probability mass function of M is assumed to be correct. Thus, to implement the sensitivity analysis, we develop a semiparametric estimator of the natural direct effect in the union model $\mathcal{M}_a \cup \mathcal{M}_c$, assuming $t(\cdot,\cdot,\cdot) = t_{\lambda^*}(\cdot,\cdot,\cdot)$ for a fixed λ^*. The proposed doubly robust estimator of the natural direct effect is then given by $\widehat{\theta}_0^{\text{doubly}}(\lambda^*) - \widehat{\delta}_0^{\text{doubly}}$ where $\widehat{\delta}_0^{\text{doubly}}$ is as previously described, and

$$\widehat{\theta}_0^{\text{doubly}}(\lambda^*) = \mathbb{P}_n\Bigg[\frac{I\{E=1\}\widehat{f}^{\text{par}}_{M|E,X}(M|E=0,X)}{\widehat{f}^{\text{par}}_{E|X}(1|X)\widehat{f}^{\text{par}}_{M|E,X}(M|E=1,X)} \times \{Y - \widehat{\mathbb{E}}^{\text{par}}(Y|X,M,E=1)\} + \widetilde{\eta}^{\text{par}}(1,0,X;\lambda^*)\Bigg],$$

with

$$\widetilde{\eta}^{\text{par}}(1,0,X;\lambda^*) = \sum_{m\in\mathcal{S}}\big\{\widehat{\mathbb{E}}^{\text{par}}(Y|X,M=m,E=1) + t_{\lambda^*}(0,m,X)\big(1 - \widehat{f}^{\text{par}}_{M|E,X}(m|E=0,X)\big) - t_{\lambda^*}(1,m,X)\big(1 - \widehat{f}^{\text{par}}_{M|E,X}(m|E=1,X)\big)\big\} \times \widehat{f}^{\text{par}}_{M|E,X}(m|E=0,X).$$

Our sensitivity analysis then entails reporting the set $\{\widehat{\theta}_0^{\text{doubly}}(\lambda) - \widehat{\delta}_0^{\text{doubly}} : \lambda\}$ (and the associated confidence intervals), which summarizes how sensitive inferences are to a deviation from the ignorability assumption $\lambda = 0$. A theoretical justification for the approach is given by the following formal result, which is proved in the supplemental Appendix.

THEOREM 4. *Suppose* $t(\cdot,\cdot,\cdot) = t_{\lambda^*}(\cdot,\cdot,\cdot)$; *then under the consistency, positivity assumptions and the ignorability assumption for the exposure,* $\widehat{\theta}_0^{\text{doubly}}(\lambda^*) - \widehat{\delta}_0^{\text{doubly}}$ *is a CAN estimator of the natural direct effect in* $\mathcal{M}_a \cup \mathcal{M}_c$.

The influence function of $\widehat{\theta}_0^{\text{doubly}}(\lambda^*)$ is provided in the Appendix, and can be used to construct a corresponding confidence interval.

It is important to note that the sensitivity analysis technique presented here differs in crucial ways from previous techniques developed by Hafeman (2008), VanderWeele (2010) and Imai, Keele and Yamamoto (2010). First, the methodology of VanderWeele (2010) postulates the existence of an unmeasured confounder U (possibly vector valued) which, when included in X, recovers the sequential ignorability assumption. The sensitivity analysis then requires specification of a sensitivity parameter encoding the effect of the unmeasured confounder on the outcome within levels of (E, X, M), and another parameter for the effect of the exposure on the density of the unmeasured confounder given (X, M). This is a daunting task which renders the approach generally impractical, except perhaps in the simple setting where it is reasonable to postulate a single binary confounder is unobserved, and one is willing to make further simplifying assumptions about the required sensitivity parameters [VanderWeele (2010)]. In comparison, the proposed approach circumvents this difficulty by concisely encoding a violation of the ignorability assumption for the mediator through the selection bias function $t_\lambda(e, m, x)$. Thus the approach makes no reference and thus is agnostic about the existence, dimension and nature of unmeasured confounders U. Furthermore, in our proposal, the ignorability violation can arise due to an unmeasured confounder of the mediator-outcome relationship that is also an effect of the exposure variable, a setting not handled by the technique of VanderWeele (2010). The method of Hafeman (2008) which is restricted to binary data, shares some of the limitations given above. Finally, in contrast with our proposed double robust approach, a coherent implementation of the sensitivity analysis techniques of Imai, Keele and Yamamoto (2010), Imai, Keele and Tingley (2010) and VanderWeele (2010) rely on correct specification of all posited models. We refer the reader to VanderWeele (2010) for further discussion of Hafeman (2008) and Imai, Keele and Yamamoto (2010).

9. Discussion. The main contribution of the current paper is a theoretically rigorous yet practically relevant semiparametric framework for making inferences about natural direct and indirect causal effects in the presence of a large number of confounding factors. Semiparametric efficiency bounds are given for the nonparametric model, and multiply robust locally efficient estimators are developed that can be used when nonparametric estimation is not possible.

Although the paper focuses on a binary exposure, we note that the extension to a polytomous exposure is trivial. In future work, we shall extend our results

for marginal effects by considering conditional natural direct and indirect effects, given a subset of pre-exposure variables [Tchetgen Tchetgen and Shpitser (2011)]. These models are particularly important in making inferences about so-called moderated mediation effects, a topic of growing interest, particularly in the field of psychology [Preacher, Rucker and Hayes (2007)]. In related work, we have recently extended our results to a survival analysis setting [Tchetgen Tchetgen (2011)].

A major limitation of the current paper is that it assumes that the mediator is measured without error, an assumption that may be unrealistic in practice and, if incorrect, may result in biased inferences about mediated effects. We note that much of the recent literature on causal mediation analysis makes a similar assumption. In future work, it will be important to build on the results derived in the current paper to appropriately account for a mis-measured mediator [Tchetgen Tchetgen and Lin (2012)].

APPENDIX

PROOF OF THEOREM 1. Let $F_{O;t} = F_{Y|M,X,E;t} F_{M|E,X;t} F_{E|X;t} F_{X;t}$ denote a one-dimensional regular parametric submodel of $\mathcal{M}_{\text{nonpar}}$, with $F_{O,0} = F_O$, and let

$$\theta_t = \theta_0(F_{O;t}) = \iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}_t(Y|E=1, M=m, X=x) \times f_{M|E,X;t}(m|E=0, X=x) f_{X;t}(x)\, d\mu(m,x).$$

The efficient influence function $S_{\theta_0}^{\text{eff,nonpar}}(\theta_0)$ is the unique random variable to satisfy the following equation:

$$\nabla_{t=0}\theta_t = \mathbb{E}\{S_{\theta_0}^{\text{eff,nonpar}}(\theta_0)U\}$$

for U the score of $F_{O;t}$ at $t=0$, and $\nabla_{t=0}$ denoting differentiation w.r.t. t at $t=0$. We observe that

$$\begin{aligned}\left.\frac{\partial\theta_t}{\partial t}\right|_{t=0} &= \iint_{\mathcal{S}\times\mathcal{X}} \nabla_{t=0}\mathbb{E}_t(Y|E=1, M=m, X=x) \\ &\qquad \times f_{M|E,X}(m|E=0, X=x) f_X(x)\, d\mu(m,x) \\ &\quad + \iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=1, M=m, X=x) \\ &\qquad \times \nabla_{t=0} f_{M|E,X;t}(m|E=0, X=x) f_X(x)\, d\mu(m,x) \\ &\quad + \iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=1, M=m, X=x) \\ &\qquad \times f_{M|E,X}(m|E=0, X=x) \nabla_{t=0} f_{X;t}(x)\, d\mu(m,x).\end{aligned}$$

Considering the first term, it is straightforward to verify that

$$\iint_{\mathcal{S}\times\mathcal{X}} \nabla_{t=0}\mathbb{E}_t(Y|E=1, M=m, X=x) f_{M|E,X}(m|E=0, X=x) f_X(x)\, d\mu(m,x)$$
$$= \mathbb{E}\Big[U \frac{I(E=1)}{f_{E|X}(E|X)}\{Y - \mathbb{E}(Y|E, M=m, X=x)\}\frac{f_{M|E,X}(M|E=0, X)}{f_{M|E,X}(M|E=1, X)}\Big].$$

Similarly, one can easily verify that

$$\iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=1, M=m, X=x)\nabla_{t=0} f_{M|E,X;t}(m|E=0, X=x) f_X(x)\, d\mu(m,x)$$
$$= \mathbb{E}\Big[U \frac{I(E=0)}{f_{E|X}(E|X)}\{\mathbb{E}(Y|E=1, M=m, X=x) - \eta(1,0,X)\}\Big],$$

and finally, one can also verify that

$$\iint_{\mathcal{S}\times\mathcal{X}} \mathbb{E}(Y|E=1, M=m, X=x) f_{M|E,X}(m|E=0, X=x)\nabla_{t=0} f_{X;t}(x)\, d\mu(m,x)$$
$$= \mathbb{E}[U\{\eta(1,0,X) - \theta_0\}].$$

Thus we obtain

$$\nabla_{t=0}\theta_t = \mathbb{E}\{S^{\text{eff,nonpar}}_{\theta_0}(\theta_0)U\}.$$

Given $S^{\text{eff,nonpar}}_{\delta e}(\delta_e)$, the results for the direct and indirect effect follow from the fact that the influence function of a difference of two functionals equals the difference of the respective influence functions. Because the model is nonparametric, there is a unique influence function for each functional, and it is efficient in the model, leading to the efficiency bound results. □

PROOF OF THEOREM 2. We begin by showing that

$$\mathbb{E}\{S^{\text{eff,nonpar}}_{\theta_0}(\theta_0; \beta^*_m, \beta^*_e, \beta^*_y)\} = 0 \tag{7}$$

under model $\mathcal{M}_{\text{union}}$. First note that $(\beta^*_y, \beta^*_m) = (\beta_y, \beta_m)$ under model $\mathcal{M}_a$. Equality (7) now follows because $\mathbb{E}^{\text{par}}(Y|X, M, E=1; \beta_y) = \mathbb{E}(Y|X, M, E=1)$ and $\eta(1,0,X;\beta_y,\beta_m) = \mathbb{E}[\{\mathbb{E}^{\text{par}}(Y|X, M, E=1; \beta_y)\}|E=0, X] = \eta(1,0,X)$:

$$\mathbb{E}\{S^{\text{eff,nonpar}}_{\theta_0}(\theta_0; \beta_m, \beta^*_e, \beta_y)\}$$
$$= \mathbb{E}\Big[\frac{I\{E=1\} f^{\text{par}}_{M|E,X}(M|E=0, X; \beta_m)}{f^{\text{par}}_{E|X}(1|X; \beta^*_e) f^{\text{par}}_{M|E,X}(M|E=1, X; \beta_m)}$$

$$
\begin{aligned}
&\times \overbrace{\mathbb{E}\{Y - \mathbb{E}^{\mathrm{par}}(Y|X, M, E=1; \beta_y)|E=1, M, X\}}^{=0}\Big] \\
&+ \mathbb{E}\Big[\frac{I(E=0)}{f^{\mathrm{par}}_{E|X}(1|X; \beta^*_e)} \\
&\qquad \times \overbrace{\mathbb{E}[\{\mathbb{E}^{\mathrm{par}}(Y|X, M, E=1; \beta_y) - \eta(1, 0, X; \beta_y, \beta_m)\}|E=0, X]}^{=0}\Big] \\
&+ \mathbb{E}[\eta(1, 0, X; \beta_y, \beta_m)] - \theta_0 \\
&= 0.
\end{aligned}
$$

Second, $(\beta^*_y, \beta^*_e) = (\beta_y, \beta_e)$ under model $\mathcal{M}_b$. Equality (7) now follows because $\mathbb{E}^{\mathrm{par}}(Y|X, M, E=1; \beta_y) = \mathbb{E}(Y|X, M, E=1)$ and $f^{\mathrm{par}}_{E|X}(1|X; \beta_e) = f_{E|X}(1|X)$:

$$
\begin{aligned}
&\mathbb{E}\{S^{\mathrm{eff,nonpar}}_{\theta_0}(\theta_0; \beta^*_m, \beta_e, \beta_y)\} \\
&\quad = \mathbb{E}\Big[\frac{I\{E=1\} f^{\mathrm{par}}_{M|E,X}(M|E=0, X; \beta^*_m)}{f^{\mathrm{par}}_{E|X}(1|X; \beta_e) f^{\mathrm{par}}_{M|E,X}(M|E=1, X; \beta^*_m)} \\
&\qquad \times \overbrace{\mathbb{E}\{Y - \mathbb{E}^{\mathrm{par}}(Y|X, M, E=1; \beta_y)|E=1, M, X\}}^{=0}\Big] \\
&\qquad + \mathbb{E}\Big[\frac{I(E=0)}{f^{\mathrm{par}}_{E|X}(1|X; \beta_e)} \\
&\qquad\qquad \times \mathbb{E}[\{\mathbb{E}^{\mathrm{par}}(Y|X, M, E=1; \beta_y) - \eta(1, 0, X; \beta_y, \beta^*_m)\}|E=0, X]\Big] \\
&\qquad + \mathbb{E}[\eta(1, 0, X; \beta_y, \beta^*_m)] - \theta_0 \\
&\quad = \mathbb{E}\big[\mathbb{E}[\{\mathbb{E}^{\mathrm{par}}(Y|X, M, E=1; \beta_y)\}|E=0, X]\big] - \theta_0 = 0.
\end{aligned}
$$

Third, equality (7) holds under model $\mathcal{M}_c$ because

$$
\begin{aligned}
&\mathbb{E}\{S^{\mathrm{eff,nonpar}}_{\theta_0}(\theta_0; \beta_m, \beta_e, \beta^*_y)\} \\
&\quad = \mathbb{E}\Big[\frac{I\{E=1\} f^{\mathrm{par}}_{M|E,X}(M|E=0, X; \beta_m)}{f^{\mathrm{par}}_{E|X}(1|X; \beta_e) f^{\mathrm{par}}_{M|E,X}(M|E=1, X; \beta_m)} \\
&\qquad \times \mathbb{E}\{Y - \mathbb{E}^{\mathrm{par}}(Y|X, M, E=1; \beta^*_y)\}\Big] \\
&\qquad + \mathbb{E}\Big[\frac{I(E=0)}{f^{\mathrm{par}}_{E|X}(1|X; \beta_e)}
\end{aligned}
$$

$$
\begin{aligned}
&\times \mathbb{E}[\{\mathbb{E}^{\text{par}}(Y|X, M, E=1; \beta_y^*) - \eta(1, 0, X; \beta_y^*, \beta_m)\}|E=0, X]\Big] \\
&+ \mathbb{E}[\eta(1, 0, X; \beta_y^*, \beta_m)] - \theta_0 \\
&= \mathbb{E}\big[\mathbb{E}[\{\mathbb{E}(Y|X, M, E=1)\}|E=0, X]\big] \\
&- \mathbb{E}\big[\mathbb{E}[\mathbb{E}^{\text{par}}(Y|X, M, E=1; \beta_y^*)|E=0, X]\big] \\
&+ \mathbb{E}\big[\mathbb{E}[\mathbb{E}^{\text{par}}(Y|X, M, E=1; \beta_y^*)|E=0, X]\big] - \mathbb{E}[\eta(1, 0, X; \beta_y^*, \beta_m)] \\
&+ \mathbb{E}[\eta(1, 0, X; \beta_y^*, \beta_m)] - \theta_0 \\
&= \mathbb{E}\big[\mathbb{E}[\{\mathbb{E}(Y|X, M, E=1)\}|E=0, X]\big] - \theta_0.
\end{aligned}
$$

Assuming that the regularity conditions of Theorem 1A in Robins, Mark and Newey (1992) hold for $S^{\text{eff,nonpar}}_{\theta_0}(\theta_0; \beta_m, \beta_e, \beta_y)$, $S_\beta(\beta)$, the expression for $S^{\text{union}}_{\theta_0}(\theta_0, \beta^*)$ follows by standard Taylor expansion arguments, and it now follows that

$$
\sqrt{n}(\widehat{\theta}_0^{\text{triply}} - \theta_0) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} S^{\text{union}}_{\theta_0, i}(\theta_0, \beta^*) + o_p(1). \tag{8}
$$

The asymptotic distribution of $\sqrt{n}(\widehat{\theta}_0^{\text{triply}} - \theta_0)$ under model $\mathcal{M}_{\text{union}}$ follows from the previous equation by Slutsky's Theorem and the Central Limit Theorem.

We note that $\widehat{\delta}_e^{\text{doubly}}$ is CAN in the union model $\mathcal{M}_{\text{union}}$ since it is CAN in the larger model where either the density for the exposure is correct, or the density of the mediator and the outcome regression are both correct and thus $\eta(e, e, X; \beta_y^*, \beta_m^*) = \mathbb{E}(Y|X, E=e)$. This gives the multiply robust result for direct and indirect effects. The asymptotic distribution of direct and indirect effect estimates then follows from similar arguments as above.

At the intersection submodel

$$
\frac{\partial \mathbb{E}\{S^{\text{eff,nonpar}}_{\theta_0}(\theta_0, \beta)\}}{\partial \beta^T} = 0
$$

hence

$$
S^{\text{union}}_{\theta_0}(\theta_0, \beta) = S^{\text{eff,nonpar}}_{\theta_0}(\theta_0, \beta).
$$

The semiparametric efficiency claim then follows for $\widehat{\theta}_0^{\text{triply}}$, and a similar argument gives the result for direct and indirect effects. □

PROOFS OF THEOREMS 3 AND 4. The proofs are given in the online Appendix. □

Acknowledgments. The authors would like to acknowledge Andrea Rotnitzky who provided invaluable comments that improved the presentation of the results given in Section 7. The authors also thank James Robins and Tyler VanderWeele for useful comments that significantly improved the presentation of this article.

SUPPLEMENTARY MATERIAL

Supplemental Appendix to Semiparametric theory for causal mediation analysis (DOI: 10.1214/12-AOS990SUPP; .pdf). The supplementary material gives the semiparametric efficiency theory for estimation of natural direct effects with a known model for the mediator density. The Appendix also gives the proof of Theorem 3 (stated in the Supplementary Appendix) and of Theorem 4.

REFERENCES

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, Edinburgh, Scotland, UK, July 30–August 5, 2005* 357–363.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.

CAO, W., TSIATIS, A. A. and DAVIDIAN, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96** 723–734. MR2538768

GOETGELUK, S., VANSTEELANDT, S. and GOETGHEBEUR, E. (2008). Estimation of controlled direct effects. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 1049–1066. MR2530329

HAFEMAN, D. (2008). Opening the black box: A reassessment of mediation from a counterfactual perspective. PhD dissertation, Columbia Univ., New York.

HAFEMAN, D. M. and VANDERWEELE, T. J. (2011). Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* **22** 753–764.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242

IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychological Methods* **15** 309–334.

IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71. MR2741814

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)* 411–442. Morgan Kaufmann, San Francisco, CA.

PEARL, J. (2011). The mediation formula: A guide to the assessment of causal pathways in nonlinear models. Technical report. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r379.pdf.

PREACHER, K. J., RUCKER, D. D. and HAYES, A. F. (2007). Assessing moderated mediation hypotheses: Strategies, methods, and prescriptions. *Multivariate Behavioral Research* **42** 185–227.

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

ROBINS, J (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (P. Green, N. Hjort and S. Richardson, eds.) 70–81. Oxford Univ. Press, Oxford.

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

ROBINS, J. M., MARK, S. D. and NEWEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48** 479–495. MR1173493

ROBINS, J. M. and RICHARDSON, T. S. (2012). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures* (P. Shrout, ed.). Oxford Univ. Press. To appear.

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

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). 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* (*Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681

ROBINS, J., SUED, M., LEI-GOMEZ, Q. and ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statist. Sci.* **22** 544–559. MR2420460

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

TAN, Z. (2010). Bounded, efficient, and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682.

TCHETGEN TCHETGEN, E. J. (2011). On causal mediation analysis with a survival outcome. *Int. J. Biostat.* **7** Art. 33, 38. MR2843528

TCHETGEN TCHETGEN, E. J. and LIN, S. H. (2012). Robust estimation of pure/natural direct effects with mediator measurement error. Technical report, Dept. Epidemiology, Harvard School of Public Health.

TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2011). Semiparametric estimation of models for natural direct and indirect effects. Harvard Univ. Biostatistics Working Paper 129. Available at http://biostats.bepress.com/harvardbiostat/paper129.

TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2012). Supplement to "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis." DOI:10.1214/12-AOS990SUPP.

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. Harvard Univ. Biostatistics Working Paper 148. Available at http://biostats.bepress.com/harvardbiostat/paper148.

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. MR2233926

VAN DER LAAN, M. and PETERSEN, M. (2005). Direct effect models. Working Paper 187. Univ. California Berkeley Division of Biostatistics Working Paper Series. Available at http://www.bepress.com/ucbbiostat/paper187.

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York. MR1958123

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

VANDERWEELE, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* **21** 540–551.

VANDERWEELE, T. J. and VANSTEELANDT, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.* **172** 1339–1348.

DEPARTMENTS OF EPIDEMIOLOGY AND BIOSTATISTICS
HARVARD SCHOOL OF PUBLIC HEALTH
677 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: etchetge@hsph.harvard.edu

DEPARTMENT OF EPIDEMIOLOGY
HARVARD SCHOOL OF PUBLIC HEALTH
677 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: ishpitse@hsph.harvard.edu