Original Article

Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis

Eric J. Tchetgen Tchetgen

Departments of Epidemiology and Biostatistics,

Harvard University

Correspondence: Eric J. Tchetgen Tchetgen, Department of Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Abstract

An important scientific goal of studies in the health and social sciences is increasingly to determine to what extent the total effect of a point exposure, treatment or intervention on a subsequent outcome is mediated by an intermediate variable on the causal pathway between the exposure and the outcome. A causal framework has recently been proposed for mediation analysis, which gives rise to new definitions, formal identification results and novel estimators of direct and indirect effects. In the present paper, the author describes a new inverse odds ratio-weighted (IORW) approach to estimate within this causal framework, so-called natural direct and indirect effects. The approach which uses as a weight, the inverse of an estimate of the odds ratio function relating the exposure to the mediator is universal in that it can be used to decompose total effects in a number of regression models commonly used in practice. Specifically, the approach may be used for effect decomposition in generalized linear models with a nonlinear link function, and in a number of other commonly used models such as the Cox proportional hazards regression for a survival outcome. The approach is simple and can be implemented in standard software provided

a weight can be specified for each observation. An additional advantage of the proposed approach is that it easily accommodates multiple mediators of a categorical, discrete or continuous nature.

KEY WORDS:

1 Introduction

Mediation analysis is an important inferential goal for many studies in the health and social sciences. In such studies, mediation analysis typically aims to quantify the extent to which a given point exposure, treatment or intervention affects the outcome of interest directly versus through an intermediate variable on the causal pathway between the exposure and the outcome. Recent developments in causal inference have provided a formalization of mediation analysis by providing counterfactual definitions, sufficient conditions for identification and a number of novel statistical methods to estimate direct and indirect effects (Robins and Greenland, 1992, Pearl, 2001, Avin et al, 2005). The current paper considers the estimation of natural direct and indirect effects (Robins and Greenland, 1992, Pearl, 2001). 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). Such an effect generally differs from the controlled direct effect which 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). The controlled direct effect combines with the controlled indirect effect to produce the joint effect of the exposure and the mediator, whereas, the natural direct and indirect effects combine to produce the exposure total effect. Pearl (2001) previously noted that 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.

Sufficient conditions for identification of natural direct and indirect effects were given by Pearl (2001, 2010); and related conditions can also be found in recent literature (Robins and Greenland,1992, Pearl, 2001, Petersen et al 2006, Hafeman and Vanderweele, 2010, Imai et al, 2010); for our purposes we shall adopt the assumptions formulated in Imai et al (2010ab) which are reproduced in Section 2. These various assumptions lead to the nonparametric identification of the natural direct and indirect effects in terms of the mediation functional of Pearl (2001,2010), which is defined in Section 2. For the purpose of estimation, previous authors have considered parametric methods (VanderWeele, 2009, VanderWeele and Vansteelandt, 2009,2010, Pearl, 2011) that posit:

(i) a model for the outcome given the exposure, mediator and pre-exposure variables,

(ii) a model for the mediator given exposure and pre-exposure variables.

and combine estimates of (i) and (ii) according to Pearl's mediation formula (2001, 2011), to form estimates of natural direct and indirect effects. Unfortunately, when conditional mediation effects are sought given covariates, models and estimates of natural direct and indirect effects, obtained using the parametric mediation formula are restricted in their functional form by the choice of models (i) and (ii). This is a potential limitation of the parametric mediation formula that is rarely discussed but nonetheless deserves some consideration. There is potentially an issue with the above approach particularly when either model (i) or model (ii) involves a non-linear link function, in which case, the parametric mediation formula induces a non-standard model of the conditional direct effect and of the conditional indirect effect; and thus of the conditional total effect. In this paper, a model for the natural direct or indirect effect, or for the total effect is considered non-standard if it does not fall within the class of regression models typically used in routine statistical applications; say a generalized linear model or a Cox proportional hazards model

for a survival outcome. To further clarify this phenomenon, suppose that a logistic regression is used in (i) to model a non-rare binary outcome, and that a logistic regression is used in (ii) to model a non-rare binary mediator, then, the parametric mediation functional combines these two standard models to produce a non-standard model of direct, indirect and total effects. Specifically, the logistic link function in (i) and (ii) dictates that the implied model for the regression of the outcome on the exposure and the covariates does not match any of the standard models typically used to estimate total effects, rendering the resulting mediation inferences difficult to interpret. An alternative to reporting conditional effects that resolves this difficulty, is to estimate marginal natural direct and indirect effects. This is the approach favored by Tchetgen Tchetgen and Shpitser (2011a) who also address concerns about possible bias due to modelling error in either (i) or (ii) and develop using modern semiparametric theory, multiply robust locally efficient estimators of marginal mean direct and indirect effects; thus extending previous similar results for total effects to the mediation context. Tchetgen Tchetgen and Shpitser (2011b) further build on this theory and propose similar multiply robust methodology to estimate parametric models for natural direct and indirect effects with an identity or log link function, conditional on a subset of pre-exposure covariates, effectively extending the work of van der Laan and Petersen (2005). Tchetgen Tchetgen (2011) further develops the semiparametric approach in marginal regression models in a survival context. Zheng and van der Laan (2011) build on the results of Tchetgen Tchetgen and Shpitser (2010ab) and obtain alternative multiply robust locally efficient targeted maximum likelihood estimators of natural direct and indirect effects on the mean difference scale.

The previous discussion sheds light on an important distinction between the parametric approach for estimating the mediation formula versus the semiparametric approach in so far that when conditional effects are sought, the latter approach directly posits a standard model for natural direct and indirect effects, and thus for the total effect, within levels of covariates; whereas

the former approach defines these effects indirectly in terms of models (i) and (ii). Despite this advantage, the semiparametric methods for conditional effects developed by Tchetgen Tchetgen and colleagues, and van der Laan and colleagues, only apply in models with an identity or log link function, and do not allow for the use of any of the other link functions often encountered in practice (e.g. logit, probit, or complementary-log link). Furthermore, semiparametric methods have not yet been developed to make inferences about mediation effects and thus to decompose conditional total effects in a Cox proportional hazards model. The main goal of this paper is to address this gap in the causal mediation literature. To achieve this goal, a new inverse odds ratio-weighted (IORW) approach is proposed for decomposing on a given scale total effects into natural direct and indirect effects. The approach which uses as a weight, an estimate of the inverse of:

(iii) the odds ratio function relating the exposure to the mediator within levels of covariates

is universal in that it can be used in a number of standard regression models commonly used to estimate total effects. Specifically, the approach may be used to decompose an exposure total effect into its direct and indirect components conditional on pre-exposure covariates, in generalized linear models with a nonlinear link function, as well as in the Cox proportional hazards model for a possibly right censored survival outcome. The approach is simple and can be implemented in standard software provided a weight can be specified for each observation. As we have indicated above, IORW estimation requires a consistent estimate of the exposure-mediator conditional odds ratio function given pre-exposure covariates. Such an estimate can be obtained by positing a working model for:

(iv) the density of the exposure given the mediator evaluated at a reference value, say zero, and pre-exposure covariates.

Together models (i) and (iv) define a model for the density of the exposure given the mediator variable and covariates which can be estimated via standard logistic regression. An advantage of this approach is that it readily scales with increasing number of mediators and thus easily accommodates multiple mediators of a categorical, discrete or continuous nature via logistic regression. A doubly robust approach is also discussed whereby working models (ii) and (iv) are combined to obtain a consistent estimate of the odds ratio function (iii) and therefore a consistent estimate of direct and indirect effects provided that the odds ratio model (iii) is correctly specified, and at least one of models (ii) or (iv) is correctly specified, but both do not necessarily hold.

2 Identification

Suppose i.i.d data on $O = (Y, E, M, X)$ is collected for n subjects, where Y denotes the 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. To formally define natural direct and indirect effects first requires defining counterfactuals. We assume for each possible level (e, m) of the exposure and mediator variables, there exist a counterfactual variable $Y_{e,m}$ corresponding to the outcome Y had possibly contrary to fact the observed exposure and mediator variables taken the value (e, m). Similarly, for $E = e$, we assume there exist a counterfactual variable M_e corresponding to the mediator variable had possibly contrary to fact the exposure variable taken the value e. To fix ideas, consider the task of decomposing on the mean scale, the conditional total

effect of E on Y given X in terms of natural direct and indirect effects :

$$\begin{aligned}
\gamma_{tot}(X) &= \overbrace{g^{-1}\{\mathbb{E}(Y_{e=1}|X)\} - g^{-1}\{\mathbb{E}(Y_{e=0}|X)\}}^{\text{total effect}} \qquad (1)\\
&= g^{-1}\{\mathbb{E}(Y_{e=1,M_{e=1}}|X)\} - g^{-1}\{\mathbb{E}(Y_{e=0,M_{e=0}}|X)\}\\
&= \overbrace{g^{-1}\{\mathbb{E}(Y_{e=1,M_{e=1}}|X)\} - g^{-1}\{\mathbb{E}(Y_{e=1,M_{e=0}}|X)\}}^{\text{natural indirect effect}} + \overbrace{g^{-1}\{\mathbb{E}(Y_{e=1,M_{e=0}}|X)\} - g^{-1}\{\mathbb{E}(Y_{e=0,M_{e=0}}|X)\}}^{\text{natural direct effect}}.\\
&= \gamma_{ind}(X) + \gamma_{dir}(X)
\end{aligned}$$

where $\mathbb{E}$ stands for expectation and g^{-1} is a user-specified nonlinear link function. The above decomposition reveals that identification of direct and indirect effects requires identification of the conditional mean of $Y_{eM_{e^*}}$ within levels of X, where $(e, e^*) \in \{0,1\}^2$. For identification, we make the following assumptions:

Consistency

$$\text{if } E = e\,, \text{then } M_e = M \text{ w.p.1},$$

$$\text{and if } E = e \text{ and } M = m \text{ then } Y_{e,m} = Y \text{ w.p.1.}$$

In addition, we adopt the sequential ignorability assumption of Imai et al (2010) which states that for $e, e^* \in \{0,1\}$:

Sequential ignorability

$$\{Y_{e^*,m}, M_e\} \perp\!\!\!\perp E|X, \qquad (2)$$

$$Y_{e^*,m} \perp\!\!\!\perp M|E = e, X, \qquad (3)$$

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

positivity

$$f_{M|E,X}(m|E,X) > 0 \text{ w.p.1 for each } m \in \mathcal{S}$$

Then, under the consistency, sequential ignorability and positivity assumptions, Imai et al (2010) showed that the cumulative distribution function (CDF) of $[Y_{eM_{e^*}}|X]$ is identified by Pearl's mediation functional :

$$F_{Y_{eM_{e^*}}|X}(y|X=x) = \int_{\mathcal{S}} F_{Y|E,M,X}(Y|E=e, M=m, X=x)\ f_{M|E,X}(m|E=e^*, X=x)\, d\mu(m) \tag{4}$$

where $F_{Y|E,M,X}$ is the CDF of $[Y|E,M,X]$ and $f_{M|E,X}$ is the conditional density of $[M|E,X]$. This in turn implies under the above assumptions, identification of various functionals of $F_{Y_{eM_{e^*}}|X}$ typically of interest; in particular, the conditional mean $\mathbb{E}(Y_{e,M_{e^*}}|X)$ is identified from the observed data; the hazard function of $[Y_{e,M_{e^*}}|X]$ is identified from the observed data when Y entails a censored failure time (provided the censoring process, and the outcome and mediator variables are independent given $(E,\ X)$).

In this paper, we chose to work under the sequential ignorability assumption of Imai et al (2010a,b) but we note that Robins and Richardson (2010) 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 made 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, i.e. they must precede E. In fact, Avin et al (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. This implies that similar 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 (2010) 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. Methods to perform such sensitivity analyses are strictly beyond the scope of the current paper, but see VanderWeele (2010), Imai et al (2010ab),Tchetgen Tchetgen and Shpitser (2011ab) and Tchetgen Tchetgen (2011) for further detail.

3 Model definition and estimation

3.1 Mediation for mean regression models

3.1.1 Estimating total effects

In this section, mediation analysis in the context of mean regression is considered. Thus, suppose that the total effect of E, $\gamma_{tot}(X)$ is estimated by fitting the mean regression model

$$g^{-1}\left(\mathbb{E}\left(Y\ |E=e,X=x;\psi\right)\right)=\widetilde{\gamma}_{tot}\left(x;\psi_{tot}\right)e+\widetilde{\gamma}^{0}\left(x;\psi_{0}\right) \tag{5}$$

where under the consistency assumption and the ignorability assumption (2),

$$\widetilde{\gamma}_{tot}\left(x;\psi_{tot}\right)=\gamma_{tot}\left(x\right)$$

is a parametric model for $\gamma_{tot}(x)$ with unknown parameter ψ_{tot}, and

$$\widetilde{\gamma}^{0}\left(x;\psi_{0}\right)=g^{-1}\left\{\mathbb{E}\left(Y_{e=0}|X\right)\right\}$$

is a parametric model for the mean of $Y_{e=0}$, with unknown parameter ψ_0; and $\psi^T=(\psi_0^T,\psi_{tot}^T)$. In practice, it is customary to specify a simple linear functional form for $\widetilde{\gamma}_{tot}$ and $\widetilde{\gamma}_{tot}^{0}$ such as for example $\left[1,x^T\right]\psi$ where ψ is of dimension $(1+\dim(X))$, but more elaborate possibly nonlinear functions of x equally apply. For estimation suppose that ψ is estimated by the vector $\widehat{\psi}$ which satisfies the empirical first order condition:

$$0=\mathbb{P}_{n}\left[\Delta_{tot}\left(E,X;\widehat{\psi}\right)\left\{Y-\mathbb{E}\left(Y\ |E=e,X=x;\widehat{\psi}\right)\right\}\right] \tag{6}$$

where $\Delta_{tot}\left(E, X; \widehat{\psi}\right)$ is a vector of size $\dim(\psi)$, and $\mathbb{P}_n[\cdot] = n^{-1}\sum_i [\cdot]_i$. A convenient choice for $\Delta_{tot}(e, x; \psi)$ is $\frac{\partial\{\widetilde{\gamma}_{tot}(x;\psi_{tot})e+\widetilde{\gamma}_0(x;\psi_0)\}}{\partial\psi^T}$, however, one should note that the maximum likelihood estimator in a generalized linear model with a mean specified by (5), typically solves a score equation of the form (6) and therefore the above class of estimating equations is quite general.

3.1.2 IORW estimation of direct effects

Now, similarly to $\widetilde{\gamma}_{tot}$, let $\widetilde{\gamma}_{dir}(x; \beta_{dir})$ denote a parametric model for $\gamma_{dir}(x)$ with unknown parameter β_{dir}. To estimate natural direct effects, we further assume that $\mathbb{E}(Y_{e,M_{e=0}}|X)$ is of the parametric form:

$$g^{-1}\left\{\mathbb{E}\left(Y_{e,M_{e=0}}|X = x; \beta_{dir}, \psi_0\right)\right\} = \widetilde{\gamma}_{dir}(x; \beta_{dir})\, e + \widetilde{\gamma}^0(x; \psi_0) \tag{7}$$

where $\beta = (\beta_{dir}, \psi_0)$. As in the model for the total effect of E, the function $\widetilde{\gamma}_{dir}$ may be specified as a simple linear function of the covariates, but more general functional forms may also be used. Let $\mathbf{OR}(M, E|X)$ denote the conditional odds ratio function relating M and E within levels of X, that is

$$\mathbf{OR}(M, E|X) = \frac{f_{M|E,X}(M|E, X)\, f_{M|E,X}(M = m_0|E = 0, X)}{f_{M|E,X}(M = m_0|E, X)\, f_{M|E,X}(M|E = 0, X)} \tag{8}$$

$$= \frac{f_{E|M,X}(E|M, X)\, f_{E|M,X}(E = 0|M = m_0, X)}{f_{E|M,X}(E = 0|M, X)\, f_{E|M,X}(E|M = m_0, X)} \tag{9}$$

where $f_{E|M,X}$ denotes the conditional density of $[E|M, X]$ and m_0 is a reference value for M. The following result motivates our estimation strategy. Before stating the result, define for any β^*, the function

$$U(\beta^*) = \mathbf{OR}(M, E|X)^{-1}\, \Delta_{dir}(E, X; \beta^*)\left\{Y - b(E, X; \beta^*)\right\}$$

where $b(e,x;\beta^*) = g\left(\widetilde{\gamma}_{dir}\left(x;\beta^*_{dir}\right)e + \widetilde{\gamma}^0\left(x;\psi^*_0\right)\right)$ and Δ_{dir} is defined similarly to Δ_{tot}.

Theorem 1: Under the consistency, sequential ignorability and positivity assumptions, and assuming model (7) *is correctly specified, we have that* $U(\beta^*)$ *is an unbiased estimating equation, in other words,* $\beta^* = \beta$ *solves the population estimating equation*

$$\mathbb{E}\left\{U\left(\beta^*\right)\right\} = 0$$

According to the theorem, estimation of β under our assumptions requires estimation of the odds ratio function $\mathbf{OR}(M, E|X)$ which is generally unknown. To proceed with inference, we assume $\mathbf{OR}(M, E|X)$ follows a parametric model $\widetilde{\mathbf{OR}}(M, E|X;\alpha_1)$ with unknown parameter α_1. Then, based on the second representation (9) of the odds ratio function, we propose to estimate $\alpha = (\alpha_0, \alpha_1)$ by fitting using maximum likelihood, the logistic regression model:

$$\text{logit}\Pr(E = 1|M = m, X = x;\alpha) = \log\widetilde{\mathbf{OR}}(m, 1|x;\alpha_1) + \log\widetilde{\mathbf{ODDS}}(x;\alpha_0) \tag{10}$$

where $\log\widetilde{\mathbf{ODDS}}(x;\alpha_0)$ is a parametric model for the baseline log odds function $\text{logit}\Pr(E = 1|M = m_0, x)$ with unknown parameter α_0. Let $\widehat{\alpha}_1$ and $\widetilde{\mathbf{OR}}(m, 1|x;\widehat{\alpha}_1)$ denote the MLEs of α_1 and $\widetilde{\mathbf{OR}}(m, 1|x;\alpha_1)$, respectively. The estimator $\widehat{\beta}$ of β then solves the equation

$$\mathbb{P}_n\left\{U\left(\widehat{\beta}, \widehat{\alpha}_1\right)\right\} = 0$$

where for all (β^*, α_1^*), $U(\beta^*, \alpha_1^*)$ is defined as $U(\beta^*)$ upon substituting $\mathbf{OR}(m, 1|x;\widehat{\alpha}_1)$ for $\mathbf{OR}(m, 1|x)$. Then, under the assumptions of Theorem 1, and the additional assumption that model (10) is correctly specified, $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ is, under sufficient regularity conditions, asymptotically normal, with

variance-covariance matrix consistently estimated by

$$\widehat{\Sigma}_1 = \widehat{\Omega}\ \widehat{\Gamma}\widehat{\Omega}^{\ T}$$

where $\widehat{\Omega}$ and $\widehat{\Gamma}$ are defined in the appendix. In practice, $\left(\widehat{\beta}, \widehat{\alpha}_1\right)$ may be obtained by fitting using standard software, a weighted generalized linear model with IOWR weight; a task easily accomplished in most software packages, e.g. by using proc genmod in SAS. For inference, it is natural to use $\widehat{\Sigma}$ to construct 95%CI for β; alternatively, the nonparametric bootstrap could be used.

3.1.3 Estimation of indirect effects

Upon obtaining $\widetilde{\gamma}_{dir}\left(x; \widehat{\beta}_{dir}\right)$ and $\widetilde{\gamma}_{tot}\left(x; \widehat{\psi}_{tot}\right)$ using the steps outlined in the previous sections, equation (5) produces the following estimator of the natural indirect effect :

$$\widetilde{\gamma}_{ind}\left(x; \widehat{\psi}_{tot}, \widehat{\beta}_{dir}\right) = \widetilde{\gamma}_{tot}\left(x; \widehat{\psi}_{tot}\right) - \widetilde{\gamma}_{dir}\left(x; \widehat{\beta}_{dir}\right)$$

with consistent variance-covariance matrix $\widehat{\Sigma}_x$ derived in the appendix.

An alternative approach

At this juncture, we should note that the above strategy for estimating γ_{dir} and γ_{ind} is asymmetric in its treatment of direct and indirect effects, and the approach clearly privileges γ_{dir} which is directly modeled while γ_{ind} is deduced from γ_{dir} and γ_{tot}. In some settings, it may be of interest to instead privilege the indirect effect by directly specifying a model $\widetilde{\gamma}_{ind}(x; \beta_{ind})$ for γ_{ind}, in which case, the counterfactual model (7) is defined in terms of γ_{ind} and γ_{tot} :

$$g^{-1}\left\{\mathbb{E}\left(Y_{e,M_{e=0}}|X = x; \psi_{tot}, \beta_{ind}, \psi_0\right)\right\} = offset\left(e, x; \psi_{tot}\right) - \widetilde{\gamma}_{ind}\left(x; \beta_{ind}\right) e + \widetilde{\gamma}^0\left(x; \psi_0\right) \qquad (11)$$

with offset:

$$offset\left(e,x;\psi_{tot}\right)=\gamma_{tot}\left(x;\psi_{tot}\right)e$$

A consistent and asymptotically normal estimator $\left(\widehat{\beta}_{ind}^{\dagger},\widehat{\psi}_{0}^{\dagger}\right)$ of (β_{ind},ψ_0) in model (11) is obtained by using the IORW approach described in the previous section upon substituting $offset\left(e,x;\widehat{\psi}_{tot}\right)$ for the unknown offset. The variance-covariance matrix of the resulting estimator $\left(\widehat{\beta}_{ind}^{\dagger},\widehat{\psi}_{0}^{\dagger}\right)$ is provided in the appendix.

3.1.4 A comparison to the parametric mediation formula

As mentioned in the introduction, the parametric mediation formula approach involves estimating a model for the mean regression of the outcome given the exposure, mediator and pre-exposure variables. To fix ideas, suppose that the following simple model is used:

$$g^{-1}\left\{\mathbb{E}\left(Y|E=e,M=m,X=x;\omega\right)\right\}=[1,e,m^{T},x^{T}]\omega \tag{12}$$

The approach also requires a model for the joint conditional density of $[M|E,X]$ which we denote $f(M|E,X;\alpha_1,\kappa)$ defined as followed:

$$f_{M|E,X}(M|E,X;\alpha_1,\kappa)=\frac{f_{M|E,X}(M|E=0,X;\kappa)\widetilde{\mathbf{OR}}\left(M,E|X;\alpha_1\right)}{\int f_{M|E,X}(m|E=0,X;\kappa)\widetilde{\mathbf{OR}}\left(m,E|X;\alpha_1\right)d\mu\left(m\right)} \tag{13}$$

so that κ parametrizes the baseline conditional density $f_{M|E,X}(M|E=0,X)$, and the equation in the above display makes explicit the dependence of the density of $[M|E,X]$ on the odds ratio function $\mathbf{OR}\left(M,E|X\right)$. Then, the parametric mediation functional (4) produces the following

expression for the counterfactual mean $\mathbb{E}\{Y_{eM_{e^*}}|X\}$:

$$\mathbb{E}\{Y_{eM_{e^*}}|X=x;\kappa,\omega,\alpha_1\}=\int_{\mathcal{S}}\mathbb{E}(Y|E=e,M=m,X=x;\omega)\,f_{M|E,X}(M=m|E,X;\alpha_1,\kappa)d\mu(m)$$

This expression in turn produces analytic expressions for the natural direct and indirect effects, and for the total effect in terms of (κ,ω,α_1) . Consider the model for the mean of $[Y|E,X]$ obtained with the formula above $\mathbb{E}\{Y|E=e,X=x;\kappa,\omega,\alpha_1\}=\mathbb{E}\{Y_{eM_e}|X=x;\kappa,\omega,\alpha_1\}$. Then, if as likely the case when either Y or M is binary, one of the models used in the formula above involves a nonlinear link function, then $\mathbb{E}\{Y|E=e,X=x;\kappa,\omega,\alpha_1\}$ will generally have a non-standard functional form, and therefore will not correspond to a regression model within the class of generalized linear models typically used to estimate total effects. We emphasize that this phenomenon can arise even if g is the identity link. For instance, if model (13) is a logistic regression modeling a binary mediator, say $\text{logit}f_{M|E,X}(M=1|E=e,X=x;\alpha_1,\kappa)=\alpha_1+[1,x^T]\kappa$, The resulting model for the mean of $[Y|E,X]$ is of the form:

$$\mathbb{E}(Y|E=e,X=x;\kappa,\omega,\alpha_1)=[1,e,\widetilde{p}(e,x;\alpha_1,\kappa),x^T]\omega$$

where $\widetilde{p}(e,x;\alpha_1,\kappa)=\left(\left\{1+\exp(-\alpha_1 e-[1,x^T]\kappa)\right\}^{-1}\right.$. Because the model in the above display will seldom be of interest in the context of total effects, mediation inferences obtained using the above modeling framework may be difficult to interpret. We should note that, there are specific settings where the above approach remains appropriate. Perhaps the most common such setting is one where both the outcome and mediator variables are continuous, and a linear regression is used to estimate their respective mean functions. Then, the parametric mediation formula is known to recover the classical Baron and Kenny (1986) approach and solely involves the mean

regression parameters (VanderWeele and Vansteelandt, 2009). Settings also exist in which in which the parametric mediation functional remains interpretable even though the outcome is a binary variable (VanderWeele and Vansteelandt, 2009, 2010); but in general, as argued above, when certain nonlinearities are present, the parametric mediation formula generally does not deliver inferences that are easily interpretable.

Furthermore, the parametric mediation approach may be particularly challenging to implement if M is multivariate (possibly with both continuous and categorical components), especially if g in (12) is nonlinear, because the approach may require modeling the joint conditional density of $[M|E=0,X]$ which may be difficult to specify correctly.

In sharp contrast, as argued throughout, IORW estimation circumvents both of the above difficulties. This is because IORW does not need a model for the density (or mean) of $[Y|E,M,X]$ or $[M|E=0,X]$, neither of which is directly of interest. In addition, as previously described, multiple mediators are easily incorporated via multiple logistic regression such as (10) . IORW is applied to the Cox proportional hazards regression in Section 3.2, but first, IORW is illustrated in a data example.

3.1.5 A data example

In this section, we conduct a mediation analysis within the context of a real world application from the psychology literature. We re-analyze data from The Job Search Intervention Study (JOBS II) also analyzed by Imai et al (2010b). 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, 1,801 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.

We consider two analyses. In the first analysis, the continuous outcome Y encodes depressive symptoms based on the Hopkins Symptom Checklist; while in the second analysis, Y is a binary variable indicating whether subjects were working more than 20 hours a week 6 months after the job training program. Both analyses consider a continuous measure of job search self-efficacy as the hypothesized mediating variable M. (Vinokur, Price, & Schul, 1995; Vinokur & Schul, 1997, Imai et al, 2010b). 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.

Continuous outcome For estimation in the context of the continuous outcome, g is set equal to the identity link, and

$$\begin{aligned} &\widetilde{\gamma}_{tot}\left(x;\psi_{tot}\right)e + \widetilde{\gamma}^{0}\left(x;\psi_{0}\right) \\ &= \psi_{tot}e + \psi_{0}x \end{aligned} \tag{14}$$

and

$$\begin{aligned} &\widetilde{\gamma}_{dir}\left(x;\beta_{dir}\right)e + \widetilde{\gamma}^{0}\left(x;\psi_{0}\right) \\ &= \beta_{dir}e + \psi_{0}x \end{aligned} \tag{15}$$

which are similar to the models estimated by Imai et al (2010b). Therefore the natural indirect effect is $\psi_{tot} - \beta_{dir}$. In addition, we set

$$\log \widetilde{\mathbf{OR}}\,(m, 1|x; \alpha_1) = \alpha_1 \tag{16}$$

$$\log \widetilde{\mathbf{ODDS}}\,(x; \alpha_0) = \alpha_0^T x \tag{17}$$

The odds ratio parameter α_1 was estimated to be $\widehat{\alpha}_1$ =0.2 (s.e.=0.08), indicating a significant difference between treatment arms in terms of job search self-efficacy. Table 1 compares results obtained using IORW estimation versus the parametric mediation formula as in Imai et al (2010b).

Insert Table 1 here.

As previously noted, the parametric mediation formula in this specific setting, coincides with the classical Baron and Kenny approach, and only requires the parameters of the following two linear regressions:

$$Y = [1, E, M, X^T]\vartheta + \epsilon_y$$

$$M = [1, E, X^T]\varphi$$

Estimates of both natural direct and indirect effects closely agreed with the results reported in Imai et al (2010b), with comparable efficiency. The results suggest a small but statistically significant mediation effect which implies that the program participation on average decreases slightly the depressive symptoms (negative average total effect) by increasing the level of job search self-efficacy.

For binary Y, we estimated conditional direct and indirect effects on the odds ratio scale (i.e. with g =logit), by using IORW. As argued in section 3.1.4, odds ratio direct and indirect

effects cannot generally be obtained using the parametric mediation formula without fairly strong distributional or related assumptions, such as a rare outcome assumption, which is known not hold in the current application. Therefore, only IORW results are reported. The results summarized in Table 1 suggests that, unlike what was observed for the depression outcome, the estimated mediation effect is small and not statistically significant, and that the estimated average total effect is larger than the estimated mediation effect, but not statistically significant.

3.2 Mediation analysis in the Cox proportional hazards model

This section concerns the decomposition of the total effect of an exposure in a Cox proportional hazards model. Thus, our goal is to estimate the natural direct and indirect effects on the hazards ratio scale:

$$\begin{aligned} HR_{tot}(x) &= \frac{\lambda_{Y_{e=1}|X}(y|X=x)}{\lambda_{Y_{e=0}|X}(y|X=x)} = \frac{\overbrace{\lambda_{Y_{e=1M_{e=1}|X}}(y|X=x)}^{\text{total effect}}}{\lambda_{Y_{e=0M_{e=0}|X}}(y|X=x)} \\ &= \frac{\overbrace{\lambda_{Y_{e=1M_{e=1}|X}}(y|X=x)}^{\text{natural indirect effect}}}{\lambda_{Y_{e=1M_{e=0}|X}}(y|X=x)} \times \frac{\lambda_{Y_{e=1M_{e=0}|X}}(y|X=x)}{\underbrace{\lambda_{Y_{e=0M_{e=0}|X}}(y|X=x)}_{\text{natural direct effect}}} \\ &= HR_{ind}(x) \times HR_{dir}(x) \end{aligned} \tag{18}$$

As before, we assume that (X, E, M) is observed on all individuals, but because of censoring, we observe $D = I(Y \leq C)$ and $Y^* = \min(Y, C)$ where C denotes an individual's right censoring time. Censoring is assumed to be independent of (Y, M) given (E, X). To proceed, suppose that

a standard Cox regression model is used to estimate the total effect of E :

$$H_{Y|E,X}(y|E=e,X=x)=H_0(y)\exp\left\{\widetilde{HR}_{tot}\left(x;\mu_{tot}\right)e+\widetilde{HR}^{0}\left(x;\mu_0\right)\right\} \tag{19}$$

where $H_{Y|E,X}$ is the hazard function of $[Y|E,X]$ and $H_0(y)$ is the hazard function of $[Y|E=0,X=0]$; thus $\widetilde{HR}_{tot}\left(x;\mu_{tot}\right)=\log HR_{tot}\left(x\right)$ is a parametric model for the total effect of E with unknown parameter μ_{tot}; $\widetilde{HR}^{0}\left(x;\mu_0\right)$ is a parametric model for the association between X and $Y_{e=0}$ on the log hazards ratio scale with restriction $\widetilde{HR}^{0}\left(0;\cdot\right)=\widetilde{HR}^{0}\left(\cdot;0\right)=0$, and μ_0 is unknown. The parameter $\left(\mu_{tot}^{T},\mu_0^{T}\right)$ may be estimated with the usual maximum partial likelihood estimator which we denote $\left(\widehat{\mu}_{tot}^{T},\widehat{\mu}_0^{T}\right)$. To estimate the hazards ratio natural direct effect HR_{dir}, we specify a Cox regression model for the counterfactual outcome Y_{eM_0} within levels of X, thus making explicit the proportional hazards assumption implicit in the effect decomposition (18):

$$H_{Y_{eM_0}|X}(y|X=x;\theta)=H_0(y)\exp\left\{\widetilde{HR}_{dir}\left(x;\theta_{dir}\right)e+\widetilde{HR}^{0}\left(x;\mu_0\right)\right\} \tag{20}$$

so that $\widetilde{HR}_{dir}\left(x;\theta_{dir}\right)=\log HR_{dir}\left(x\right)$ is a model that encodes the direct effect of E with unknown parameter θ_{dir}, and $\theta=\left(\theta_{dir},\mu_0\right)$.The following result motivates our strategy for estimating θ_{dir}. Before stating the result, define for any θ^{*}, the estimating function

$$U_{ph}\left(\theta^{*}\right)=\int dN^{*}(y)\mathbf{OR}\left(M,E|X\right)^{-1}\left[\Delta_{ph}\left(E,X;\theta^{*}\right)-\frac{\xi_1\left(y;\theta^{*}\right)}{\xi_0\left(y;\theta^{*}\right)}\right],$$

where

$$\xi_j\left(y;\theta^{*}\right)=\mathbb{E}\left[\mathbf{OR}\left(M,E|X\right)^{-1}\Delta_{ph}\left(E,X;\theta^{*}\right)^{j}\exp\left\{\widetilde{HR}_{dir}\left(x;\theta_{dir}^{*}\right)E+\widetilde{HR}^{0}\left(X;\mu_0^{*}\right)\right\}R(y)\right],$$

$N^*(y) = I(\min(Y,C) \leq y, D = 1)$ is the counting process of an observed failure time, $R(y) = I(\min(Y,C) > y)$ is the at-risk process, and:

$$\Delta_{ph}(e,x;\theta^*) = \frac{\left\{\widetilde{HR}_{dir}(x;\theta^*_{dir})\, e + \widetilde{HR}^0(x;\mu^*_0)\right\}}{\partial\theta^*}$$

$U_{ph}(\theta^*)$ is of the form of a weighted version of the score function of the partial likelihood in a Cox proportional hazards model. The next theorem states that the inverse odds ratio weight is key to identifying direct effects on a hazards ratio scale.

Theorem 2: Under the consistency, sequential ignorability and positivity assumptions, and assuming model (20) *is correctly specified and censoring is independent of* M *given* (E,X), *we have that* $U_{ph}(\theta^*)$ *is an unbiased estimating equation, in other words,* $\theta^* = \theta$ *solves the population estimating equation*

$$\mathbb{E}\{U_{ph}(\theta^*)\} = 0$$

A feasible estimating equation is obtained by replacing unknown expectations with their empirical version, and upon substituting $\widetilde{\mathbf{OR}}(M,E|X;\widehat{\alpha}_1)$ for the unknown weight $\mathbf{OR}(M,E|X)$. The resulting estimator of θ is, under the assumptions of theorem 2, and the additional assumption that model (10) is correctly specified, consistent and asymptotically normal under standard regularity conditions. For inference, we recommend using the nonparametric bootstrap.

We should note that the estimator described in the previous paragraph can easily be obtained using standard Cox regression software, such as proc phreg in SAS, which provides an option for user-specified weights. Natural indirect effect estimates naturally follow from the relation (18). Our exposition has again given priority to natural direct effects over indirect effects in the sense that a model is chosen for the latter in terms of models for direct and total effects. Similarly

to mean models, it is possible to prioritize the indirect effect and the total effect and to express models for the direct effect in terms of these models. Details for estimation are omitted but are easily deduced from the exposition.

3.2.1 A data example

We briefly illustrate the methods described in this section with a reanalysis of a study by Caplehorn and Bell (1991) which compares two methadone treatment clinics for heroin addicts to assess patient time remaining under methadone treatment. A patient's survival time was determined as the time, in days until the person dropped out of the clinic or was censored. The two clinics differed according to their live-in policies for patients. Here we wish to infer the degree to which patients' methadone dosage mediates differences in retention of patients in the two clinics. In addition to the exposure (E = indicator of which methadone treatment clinic the patient attended), mediator (M = a continuous variable for the patient's maximum methadone dose (mg/day)) and outcome (Y = time until the patient dropped out of the clinic or was censored), a covariate is also available (X = indicates whether the patient had a prison record). Note that the continuous mediator is easily incorporated in the logistic regression (10) . For estimation, we used

$$\widetilde{HR}_{tot}\left(x;\mu_{tot}\right)e+\widetilde{HR}^{0}\left(x;\mu_{0}\right)=\mu_{tot}e+\mu_{0}^{T}x$$

$$\widetilde{HR}_{dir}\left(x;\theta_{dir}\right)e+\widetilde{HR}^{0}\left(x;\mu_{0}\right)=\theta_{dir}e+\mu_{0}^{T}x$$

therefore the natural indirect effect is $\mu_{tot}-\theta_{dir}$, and we estimated models (16) and (17) via logistic regression maximum likelihood. The odds ratio parameter α_1 was estimated to be $\widehat{\alpha}_1$ =-0.02 (s.e.=0.009), indicating a significant difference between clinics in terms of patient's methadone dosage.

Insert Table 2 here.

Table 2 summarizes results based on 266 patients included in the Caplehorn study (37% of whom were censored). The analysis establishes the presence of a large clinic total effect on the hazards ratio scale, and suggests that most of this effect is not mediated by methadone dose and is direct. Nonetheless, these results should be interpreted with caution and should only be taken as an illustrative example of the methodology, because it may not be realistic to assume that X contains all patients baseline correlates of E, M and Y (beyond a prison record) in this data set, as required for sequential ignorability to hold.

3.3 Doubly robust estimation

Throughout, we have assumed that $\widetilde{\mathbf{OR}}\,(m, 1|x; \widehat{\alpha}_1)$ consistently estimates $\mathbf{OR}\,(M, E|X)$, which requires that models $\widetilde{\mathbf{OR}}$ and $\widetilde{\mathbf{ODDS}}$ are both correct. Modeling error of either of these models will in general produce biased and therefore erroneous mediation inferences about the effects of E. Here we propose to increase the robustness of the proposed methodology when $\widetilde{\mathbf{OR}}$ is correctly specified. To do so, we propose to use, the doubly robust estimator of odds ratios proposed by Tchetgen Tchetgen and colleagues (Tchetgen Tchetgen et al, 2010). In addition to $\widetilde{\mathbf{ODDS}}$, the doubly robust approach also uses an estimate of the working model $f_{M|E,X}(M|E = 0, X; \kappa)$ of the density of M in the unexposed, within levels of X. However, the doubly robust approach produces a consistent and asymptotically normal estimate of $\widetilde{\mathbf{OR}}$ provided that at least one of $\widetilde{\mathbf{ODDS}}$ or $f_{M|E,X}(M|E = 0, X; \kappa)$ is correctly specified, but both models do not necessarily need to hold.

For brevity, suppose that M is binary and let

$$\log \widetilde{\mathbf{OR}}\,(M, E|X; \alpha_1) = \alpha_1 ME$$

and let $\widehat{\kappa}$ denote an estimator of κ, say the MLE under model (13). Let $W = w(X)$ be a user-specified function of X. Then, by a result due to Tchetgen Tchetgen et al (2010) , it is possible to show that $\widehat{\alpha_1}(w)$ is doubly robust and converges to α_1 provided that either $\widehat{q}_M = f_{M|E,X}(M|E = 0, X; \widehat{\kappa})$ is consistent, or $\widehat{q}_E = \left\{1 + \widetilde{\mathbf{ODDS}}(X; \widehat{\alpha}_0)^{-1}\right\}^{-1} = \Pr(E = 1|M = 0, X; \widehat{\alpha}_0)$ is consistent, where $\widehat{\alpha_1}(w) =:$

$$\log \frac{\mathbb{P}_n W \{EM(1-\widehat{q}_M)(1-\widehat{q}_E)\}}{\mathbb{P}_n [W \{M(1-E)(1-\widehat{q}_M)\widehat{q}_E + (1-M)E\widehat{q}_M(1-\widehat{q}_E) - (1-E)(1-M)(1-\widehat{q}_E)(1-\widehat{q}_M)\}]}$$

In practice, the choice $W = 1$ is convenient; the optimal choice of W can be obtained from a result due to Tchetgen Tchetgen et al (2010). The doubly robust methodology generalizes to polytomous and continuous possibly vector valued M and E, and similar methodology is available for more general models $\widetilde{\mathbf{OR}}(M, E|X; \alpha_1)$; although closed-form estimators are generally not available in such more general settings and one must resort to the methodology detailed in Tchetgen Tchetgen et al (2010).

4 Conclusion

The main contribution of the present paper is to present a simple yet general framework for making inferences about conditional natural direct and indirect causal effects that can be used to decompose total effects estimated in regression models commonly encountered in practice. The proposed IORW approach involves inverse odds ratio weights that relate exposure and mediator variables and therefore can be implemented in most standard regression software, provided that a weight can be specified. An important limitation of the proposed approach is that, similar to existing causal mediation methods, it is assumed that the mediator is measured without error. In

future work, it will be crucial to examine the extent to which a violation of this assumption might alter mediation inferences and to develop alternative methodology to appropriately account for possible measurement error of the mediator.

References

[1] Avin, C., I. Shpitser, and J. Pearl (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, pp. 357–363.

[2] Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173–1182.

[3] Caplehorn JRM, Bell J.Methadone dosage and retention of patients in maintenance treatment. Med J Australia 1991;154:195–199.

[4] Goetgeluk, S., Vansteelandt, S. and Goetghebeur, E. (2008). Estimation of controlled direct effects. Journal of the Royal Statistical Society – Series B, 70, 1049-1066.

[5] Hafeman, D. and T. VanderWeele (2009). Alternative assumptions for the identification of direct and indirect effects. Epidemiology. In press.

[6] van der Laan, M, Petersen, M. (2005) Direct Effect Models. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 187. http://www.bepress.com/ucbbiostat/paper187

[7] Imai, K., Keele, L., and Yamamoto, T. (2010a). Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science 25, 51–71.

[8] Imai, K, Keele L and Tingley D. (2010b). "A General Approach to Causal Mediation Analysis." Psychological Methods, Vol. 15, No. 4 (December), pp. 309-334.

[9] Pearl, J. (2001). Direct and indirect effects. In Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01), San Francisco, CA, pp. 411–42. Morgan Kaufmann.

[10] Pearl J (2011) The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models.Technical report <http://ftp.cs.ucla.edu/pub/stat_ser/r379.pdf>.

[11] Petersen M, Sinisi, S; van der Laan, M.(2006) Estimation of Direct Causal Effects.Epidemiology. Volume 17, 3, 276-284.

[12] Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. Epidemiology 3, 143–155.

[13] Robins, J. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. Green, N. Hjort, and S. Richardson (Eds.), Highly Structured Stochastic Systems, pp. 70–81. Oxford, UK: Oxford University Press.

[14] Robins JM, Richardson, TS. (2010). Alternative graphical causal models and the identification of direct effects. To appear in Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures. P. Shrout, Editor. Oxford University Press

[15] Tchetgen Tchetgen EJ and Shpitser I (2011) Semiparametric Theory for Causal Mediation Analysis: efficiency bounds, multiple robustness, and sensitivity analysis. Submitted for pub-

lication. Available at Harvard University Biostatistics Working Paper Series. Working Paper 130 http://www.bepress.com/harvardbiostat/paper130.

[16] Tchetgen Tchetgen EJ and Shpitser I (2011) Semiparametric Estimation of Models for Natural Direct and Indirect Effects. Submitted for publication. Available at Harvard University Biostatistics Working Paper Series. Working Paper 129.http://www.bepress.com/harvardbiostat/paper129 .

[17] Tchetgen Tchetgen EJ, Rotnitzky A and Robins J (2010). On doubly robust estimation in a semiparametric odds ratio model. Biometrika. 97 (1): 171-180.

[18] Tchetgen Tchetgen, Eric J. (2011) On Causal Mediation Analysis with a Survival Outcome, The International Journal of Biostatistics: Vol. 7: Iss. 1, Article 33.

[19] VanderWeele, T.J. (2009). Marginal structural models for the estimation of direct and indirect effects. Epidemiology, 20:18-26.

[20] VanderWeele, T.J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. Statistics and its Interface, 2, 457-468.

[21] VanderWeele, T.J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome - with discussion. American Journal of Epidemiology, 172, 1339-1348.

[22] VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology. 2010;21:540–551.

Table 1. Estimated causal effects for a continuous outcome and a binary outcome using the JOBS II data

Average effect	Continuous Y		Binary Y
	identity link		logit link
	IORW*	BK/I**	IORW
ψ_{tot}	-0.047 (0.031)	-0.047 (0.036)	0.279 (0.161)
β_{dir}	-0.033 (0.031)	-0.032 (0.039)	0.280 (0.160)
$\psi_{tot} - \beta_{dir}$	-0.014 (0.006)	-0.016(0.007)	-0.001 (0.026)

*inverse-odds ratio estimate (nonparametric bootstrap standard error)

**Baron and Kenny/Imai estimate (nonparametric boostrap standard error)

Table 2. Estimated causal effects for a failure time outcome in the Methadone example

Log(hazards ratio)	ψ_{tot}	β_{dir}	$\psi_{tot} - \beta_{dir}$
IORW*	1.10 (0.23)	1.07 (0.24)	0.07 (0.09)

*inverse-odds ratio estimate (nonparametric bootstrap standard error)

APPENDIX

Definition of $\widehat{\Omega}$ and $\widehat{\Gamma}$

$$\widehat{\Omega} = \mathbb{P}_n \left\{ \frac{\partial U\left(\beta^*, \widehat{\alpha}_1\right)}{\partial \beta^*} |_{\widehat{\beta}} \right\}^{-1},$$

$$\widehat{\Gamma} = \mathbb{P}_n \left[U\left(\widehat{\beta}, \widehat{\alpha}_1\right) + \mathbb{P}_n \left\{ \frac{\partial U\left(\beta^*, \alpha_1^*\right)}{\partial \alpha^{*T}} |_{\widehat{\alpha}_1} \right\} \mathbb{P}_n \left[S_{\alpha_1}^{eff}\left(\widehat{\alpha}\right) S_{\alpha_1}^{eff}\left(\widehat{\alpha}\right)^T \right]^{-1} S_{\alpha_1}^{eff}\left(\widehat{\alpha}\right) \right]^{\otimes 2},$$

where $v^{\otimes 2} = vv^T$, and

$$S_{\alpha_1}^{eff}\left(\widehat{\alpha}\right) = S_{\alpha_1}\left(\widehat{\alpha}\right) - \mathbb{P}_n \left[S_{\alpha_1}\left(\widehat{\alpha}\right) S_{\alpha_0}^T\left(\widehat{\alpha}\right) \right] \mathbb{P}_n \left[S_{\alpha_0}\left(\widehat{\alpha}\right) S_{\alpha_0}^T\left(\widehat{\alpha}\right) \right]^{-1} S_{\alpha_0}\left(\widehat{\alpha}\right)$$

where $\left(S_{\alpha_0}^T\left(\widehat{\alpha}\right), S_{\alpha_1}^T\left(\widehat{\alpha}\right)\right)^T$ is the score function of α in model (10).

Definition of $\widehat{\Sigma}_x$

Let $\widehat{W} = \left(\widehat{W}_1^T, \widehat{W}_2^T\right)$ where

$$\widehat{W}_1 = -\mathbb{P}_n \left[\Delta_{tot}\left(E, X; \widehat{\psi}\right) \frac{\partial \mathbb{E}\left(Y \mid E = e, X = x; \psi^*\right)}{\partial \psi^{*T}} |_{\widehat{\psi}} \right]^{-1} \Delta_{tot}\left(E, X; \widehat{\psi}\right) \left\{ Y - \mathbb{E}\left(Y \mid E = e, X = x; \widehat{\psi}\right) \right\}$$

$$\widehat{W}_2 = \widehat{\Omega} \times \left[U\left(\widehat{\beta}, \widehat{\alpha}_1\right) + \mathbb{P}_n \left\{ \frac{\partial U\left(\beta^*, \alpha_1^*\right)}{\partial \alpha^{*T}} |_{\widehat{\alpha}_1} \right\} \mathbb{P}_n \left[S_{\alpha_1}^{eff}\left(\widehat{\alpha}\right) S_{\alpha_1}^{eff}\left(\widehat{\alpha}\right)^T \right]^{-1} S_{\alpha_1}^{eff}\left(\widehat{\alpha}\right) \right]$$

and define

$$\widehat{\Theta} = \mathbb{P}_n \left\{ \widehat{W}\widehat{W}^T \right\}$$

then a consistent estimator of the variance-covariance matrix of $\gamma_{ind}\left(x; \widehat{\psi}_{tot}, \widehat{\beta}_{dir}\right)$ is obtained by a straightforward application of the delta method which yields

$$\widehat{\Sigma}_x = \dot{\gamma}_{ind}\left(x; \widehat{\psi}_{tot}, \widehat{\beta}_{dir}\right)^T \widehat{\Theta} \dot{\gamma}_{ind}\left(x; \widehat{\psi}_{tot}, \widehat{\beta}_{dir}\right)$$

where

$$\dot{\gamma}_{ind}\left(x;\widehat{\psi}_{tot},\widehat{\beta}_{dir}\right)=\left[\frac{\partial\gamma_{ind}\left(x;\psi_{tot}^{*},\beta_{dir}^{*}\right)}{\partial\left(\psi_{tot}^{T*},\beta_{dir}^{T*}\right)}|_{\left(\widehat{\psi}_{tot}^{T},\widehat{\beta}_{dir}^{T}\right)^{T}}\right]$$

Variance-covariance of $\widehat{\beta}^{\dagger}=\left(\widehat{\beta}_{ind}^{\dagger},\widehat{\psi}_{0}^{\dagger}\right)$ **:**

Define $U\left(\alpha^{*},\psi_{tot}^{*},\ \beta^{\dagger *}\right)$ as $U\left(\alpha^{*},\ \beta^{*}\right)=U\left(\alpha^{*},\ \beta_{dir}^{*},\beta_{0}^{*}\right)$ but replacing $\mathbb{E}\left(Y_{e,M_{e=0}}|X=x;\ \ \beta^{*}\right)$ with $\left\{\mathbb{E}\left(Y_{e,M_{e=0}}|X=x;\psi_{tot},\beta^{\dagger *}=\left(\beta_{ind}^{\dagger},\psi_{0}^{\dagger}\right)\right)\right\}$. Then let:

$$\begin{aligned}\widehat{W^{\dagger}}_{2}^{T}=\mathbb{P}_{n}&\left\{\frac{\partial U\left(\widehat{\alpha},\widehat{\psi}_{tot},\beta^{\dagger *}\right)}{\partial\beta^{\dagger *T}}|_{\widehat{\beta}^{\dagger}}\right\}^{-1}\times\left[U\left(\widehat{\alpha},\widehat{\psi}_{tot},\widehat{\beta}^{\dagger}\right)\right.\\&+\mathbb{P}_{n}\left\{\frac{\partial U\left(\alpha^{*},\widehat{\psi}_{tot},\widehat{\beta}^{\dagger}\right)}{\partial\alpha^{*T}}|_{\widehat{\alpha}_{1}}\right\}\mathbb{P}_{n}\left[S_{\alpha_{1}}^{eff}\left(\widehat{\alpha}\right)S_{\alpha_{1}}^{eff}\left(\widehat{\alpha}\right)^{T}\right]^{-1}S_{\alpha_{1}}^{eff}\left(\widehat{\alpha}\right)\\&\left.+\mathbb{P}_{n}\left\{\frac{\partial U\left(\widehat{\alpha},\psi_{tot}^{*},\widehat{\beta}^{\dagger}\right)}{\partial\psi_{tot}^{*T}}|_{\widehat{\psi}_{tot}}\right\}\widehat{W_{1}}\ \right]\end{aligned}$$

Then, the variance-covariance matrix of the limiting distribution of $\sqrt{n}\left(\widehat{\beta}_{ind}^{\dagger T}-\beta_{ind}^{\dagger T},\widehat{\beta}_{0}^{\dagger T}-\beta_{0}^{\dagger T}\right)$ is consistently estimated by $\mathbb{P}_{n}\left\{\widehat{W^{\dagger}}_{2}\widehat{W^{\dagger}}_{2}^{T}\right\}$.

Proof of Theorem 1

Under the consistency, sequential ignorability and positivity assumptions, and assuming model (7) is correctly specified, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{OR}\left(M,E|X\right)^{-1}\Delta_{dir}\left(E,X;\beta^{*}\right)\left\{Y-b(E,X;\beta^{*})\right\}|E,X\right]\\
&=\Delta_{dir}\left(E,X;\beta^{*}\right)\mathbb{E}\left[\frac{f_{M|E,X}\left(M=m_{0}|E,X\right)f_{M|E,X}\left(M|E=0,X\right)}{f_{M|E,X}\left(M|E,X\right)f_{M|E,X}\left(M=m_{0}|E=0,X\right)}\right.\\
&\left.\left\{Y-b(E,X;\beta^{*})\right\}|E,X\right]\\
&=\Delta_{dir}\left(E,X;\beta^{*}\right)\frac{f_{M|E,X}\left(M=m_{0}|E,X\right)}{f_{M|E,X}\left(M=m_{0}|E=0,X\right)}\\
&\times\left[\int f_{M|E,X}\left(m|E=0,X\right)\mathbb{E}\left(Y|E,M=m,X\right)d\mu\left(m\right)-b(E,X;\beta^{*})\right]\\
&=0
\end{aligned}
$$

Proof of Theorem 1

Under the consistency, sequential ignorability and positivity assumptions, and assuming model (7) is correctly specified, we have for any function $L=l(E,X)$

$$
\begin{aligned}
&\mathbb{E}\left\{\mathbf{OR}\left(M,E|X\right)^{-1}l(E,X,y)dN^{*}(y)\right\}\\
&=\mathbb{E}\left[\frac{f_{M|E,X}\left(M=m_{0}|E,X\right)}{f_{M|E,X}\left(M=m_{0}|E=0,X\right)}l(E,X)H_{Y_{E,M_{e=0}}|E,X}\left(y|E,X\right)S_{Y_{E,M_{e=0}}|E,X}\left(y|E,X\right)\right.\\
&\left.\times S_{C|E,M,X}\left(y|E,X\right)dy\right]\\
&=\mathbb{E}\left[\frac{f_{M|E,X}\left(M=m_{0}|E,X\right)}{f_{M|E,X}\left(M=m_{0}|E=0,X\right)}l(E,X)H_{Y_{E,M_{e=0}}|E,X}\left(y|E,X\right)S_{Y_{E,M_{e=0}}|E,X}\left(y|E,X\right)\right.\\
&\left.\times S_{C|E,M,X}\left(y|E,X\right)dy\right]
\end{aligned}
$$

and

$$\mathbb{E}\left\{\mathbf{OR}\left(M,E|X\right)^{-1}l(E,X)R(y)\right\}$$
$$=\mathbb{E}\left[\int\frac{f_{M|E,X}\left(M=m_0|E,X\right)}{f_{M|E,X}\left(M=m_0|E=0,X\right)}l(E,X)S_{Y_{E,M_{e=0}}|E,X}\left(y|E,X\right)$$
$$S_{C|E,M,X}\left(y|E,X\right)d\mu\left(m\right)\Big]$$

It is then straightforward using the above to establish the result, by noting that

$$\xi_j\left(y;\theta^*\right)=\mathbb{E}\left[\mathbf{OR}\left(M,E|X\right)^{-1}\Delta_{ph}\left(E,X;\theta^*\right)^j\exp\left\{\widetilde{HR}_{dir}\left(X;\theta^*_{dir}\right)E+\widetilde{HR}^0\left(X;\mu_0^*\right)\right\}R(y)\right]$$
$$=\int\int\mathbb{E}\left[\frac{f_{M|E,X}\left(M=m_0|E,X\right)}{f_{M|E,X}\left(M=m_0|E=0,X\right)}\Delta_{ph}\left(E,X;\theta^*\right)^j\exp\left\{\widetilde{HR}_{dir}\left(X;\theta^*_{dir}\right)E+\widetilde{HR}^0\left(X;\mu_0^*\right)\right\}\right.$$
$$\left.S_{Y_{E,M_{e=0}}|E,X}\left(y|E,X\right)S_{C|E,M,X}\left(y|E,X\right)\right]$$

and

$$\mathbb{E}\left[\mathbf{OR}\left(M,E|X\right)^{-1}\Delta_{ph}\left(E,X;\theta^*\right)\ dN^*(y)\right]$$
$$=\mathbb{E}\left[\frac{f_{M|E,X}\left(M=m_0|E,X\right)}{f_{M|E,X}\left(M=m_0|E=0,X\right)}\Delta_{ph}\left(E,X;\theta^*\right)\exp\left\{\widetilde{HR}_{dir}\left(X;\theta^*_{dir}\right)E+\widetilde{HR}^0\left(X;\mu_0^*\right)\right\}\right.$$
$$\left.\times S_{C|E,M,X}\left(y|E,X\right)S_{Y_{E,M_{e=0}}|E,X}\left(y|E,X\right)dy\right]\times H_{Y_{E,M_{e=0}}|E,X}\left(y|E=0,X=0\right)$$