JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **29**: 920–943 (2014)
Published online 30 June 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2341

IDENTIFYING CAUSAL MECHANISMS (PRIMARILY) BASED ON INVERSE PROBABILITY WEIGHTING

MARTIN HUBER*
Department of Economics, University of St Gallen, Switzerland

SUMMARY
This paper demonstrates the identification of causal mechanisms of a binary treatment under selection on observables, (primarily) based on inverse probability weighting; i.e. we consider the average indirect effect of the treatment, which operates through an intermediate variable (or mediator) that is situated on the causal path between the treatment and the outcome, as well as the (unmediated) direct effect. Even under random treatment assignment, subsequent selection into the mediator is generally non-random such that causal mechanisms are only identified when controlling for confounders of the mediator and the outcome. To tackle this issue, units are weighted by the inverse of their conditional treatment propensity given the mediator and observed confounders. We show that the form and applicability of weighting depend on whether some confounders are themselves influenced by the treatment or not. A simulation study gives the intuition for these results and an empirical application to the direct and indirect health effects (through employment) of the US Job Corps program is also provided. Copyright © 2013 John Wiley & Sons, Ltd.

Received 5 July 2012; Revised 16 March 2013;

Supporting information may be found in the online version of this article.

1. INTRODUCTION

A vast literature in economics and other social sciences is concerned with the evaluation of average treatment effects (ATE), both in randomized experiments and in observational studies. In many evaluations not only is the (total) ATE of interest, but also the causal mechanisms through which it operates. In this case, one would like to disentangle the *direct* effect of the treatment on the outcome as well as the *indirect* ones that run through one or more intermediate variables—so-called mediators. For example, when assessing the employment or earnings effects of an active labor market policy, policymakers might want to know to what extent the total impact comes from increased search effort, human capital or othermediators that are themselves affected by the policy. However, even in experiments, causal mechanisms are not easily identified. As discussed in Robins and Greenland (1992), random treatment assignment does not imply randomness of the mediator, which may be regarded as an intermediate outcome. Therefore, the total effect cannot be disentangled by simply conditioning on a mediator, because this generally introduces selection bias coming from variables influencing both the mediator and the outcome (see Rosenbaum, 1984).[1]

The main contribution of this paper is to show that an easily implemented version of inverse probability weighting (IPW)[2] identifies causal mechanisms under discrete or continuous mediators,

* Correspondence to: Martin Huber, University of St Gallen, Varnbüelstrasse 14, CH-9000 St Gallen, Switzerland.
E-mail: martin.huber@unisg.ch

[1] For this reason, the early work on mediation analysis of Judd and Kenny (1981) highlights the importance of controlling for such confounders. It therefore seems surprising that this issue has been ignored in so many applications in social sciences that claim to identify direct and indirect effects.

[2] The idea of IPW goes back to Horvitz and Thompson (1952), who first proposed an estimator of the population mean in the presence of non-randomly missing data.

given that a sequential selection on observables (or conditional independence) assumption holds. The latter requires (i) that the treatment is either random or exogenous given the covariates and (ii) that the mediator is exogenous given the covariates and the treatment (see, for instance, Imai *et al*., 2010). Direct and indirect effects can then be identified by weighting observations by their inverse conditional propensities to be in a particular treatment state, given (i) the observed covariates and (ii) the mediator and the observed covariates. Furthermore, we also report that the identification results for the indirect effects change if some of the covariates are themselves a function of the treatment. If the latter is the case, the identification of the 'total' indirect effect, which also accounts for correlations between those covariates affected by the treatment and the mediator, requires additional restrictions (see Robins, 2003; Avin *et al*., 2005; and Imai and Yamamoto, 2013). In contrast, the 'partial' indirect effect, which only considers the immediate link between the treatment and the mediator (and no 'detour' via any covariates), is identified under weaker assumptions. We provide a simulation study that gives the intuition for these identification issues and apply our methods to experimental data on Job Corps, a US educational program for disadvantaged youths.

The evaluation of direct and indirect effects, often referred to as mediation analysis, is widespread in social sciences such as epidemiology, political sciences and psychology (see MacKinnon, 2008). While many studies follow Baron and Kenny (1986) and rely on linear mediation models, more general identification under conditional exogeneity of the treatment and the mediator has been considered by Pearl (2001), Robins (2003), Petersen *et al*. (2006), VanderWeele (2009), Imai *et al*. (2010), Albert and Nelson (2011) and Imai and Yamamoto (2013), among others. One of the rare studies in economics is Flores and Flores-Lagunes (2009), who evaluate the direct earnings effect of Job Corps when controlling for the mediator 'work experience'. The issue is that participating in training likely decreases work experience shortly after program start compared to non-participation due to decreased job search effort during training participation ('locking-in effect'). Assuming mediator exogeneity conditional on pre-treatment covariates, Flores and Flores-Lagunes (2009) estimate a positive direct effect on earnings based on a regression approach.[3]

If conditional exogeneity does not hold (and plausible instruments are not available), point identification is lost, but partial identification based on deriving upper and lower bounds on the direct and indirect effects might still be useful. For example, Kaufman *et al*. (2005), Cai *et al*. (2008) and Sjölander (2009) focus on partial identification in randomized medical trials with binary treatments and impose specific restrictions such as monotonicity of the mediator in thetreatment in order to tighten the bounds. In economics, Flores and Flores-Lagunes (2010) consider additional assumptions (e.g. a particular order of the mean potential outcomes of various subpopulations) and assess the effectiveness of various components of the Job Corps program.

This paper focusses on point identification and makes four contributions to the literature on causal mechanisms in economics. Firstly, it derives identification results based on IPW by the treatment propensity score that are straightforward to implement by semi- or nonparametric estimation. If the mediator is exogenous conditional on pre-treatment covariates, our approach allows relaxing one functional form assumption imposed in Flores and Flores-Lagunes (2009) (their Assumption 3). It is also easier to implement than the nonparametric estimators of Imai *et al*. (2010), which require estimating the conditional mean of the outcome and the conditional density of the mediator. Secondly, and in contrast to Flores and Flores-Lagunes (2009), we also discuss identification when mediator exogeneity only holds conditional on post-treatment covariates which are themselves a function of the treatment, such that pre-treatment variables do not fully capture mediator selection. This appears realistic in most applications, including Job Corps, where the treatment likely affects variables that

[3] As a further example, Simonsen and Skipper (2006) use a semiparametric identification strategy based on matching to assess the direct wage effect of motherhood in Denmark by controlling for several mediators through which motherhood may have an influence on wages. They find negative direct effects which vary little across different sectors.

DOI: 10.1002/jae

potentially confound the mediator and the outcome, e.g. intermediate health shortly prior to the mediator. While direct effects are still identified by IPW in this set-up after a modification of the initial assumptions, the identification of indirect effects requires additional restrictions. We present a functional form restriction allowing us to do so, which, however, is less general than the entirely nonparametric identification under IPW. Thirdly, we show that IPW still identifies a partial indirect effect when keeping the confounders fixed, i.e. the part of the indirect effect not working through post-treatment confounders. Fourthly, as an empirical contribution, the present work appears to be the first which assesses the direct and indirect health effects of the Job Corps program.

The remainder of this paper is organized as follows. Section 2 defines the parameters of interest (the average direct and indirect effects) and shows identification (mostly) based on IPW. Section 3 briefly discusses estimation. Section 4 presents a simulation study which provides the intuition for various identification issues. In Section 5, we apply our methods to the experimental study of the Job Corps program. Section 6 concludes.

2. PARAMETERS OF INTEREST AND IDENTIFICATION

2.1. Definition of Parameters

Suppose we are interested in the average treatment effect (ATE) of a binary treatment indicator D on some outcome variable Y. Furthermore, assume that we would like to disentangle the ATE into a direct component and an indirect effect operating through the mediator M which has bounded support and may be discrete or continuous. To define the parameters of interest, we use the potential outcome framework advocated by Rubin (1974) (among many others) and considered in the direct and indirect effects framework, for instance, by Rubin (2004), Ten Have *et al*. (2007) and Albert (2008). Let $Y(d), M(d)$ denote the potential outcome and the potential mediator state under treatment $d \in \{0,1\}$. For each unit only one of the two potential outcomes and mediator states, respectively, is observed, because the realized outcome and mediator values are $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$ and $M = D \cdot M(1) + (1 - D) \cdot M(0)$.

The ATE is defined by $\Delta = E[Y(1) - Y(0)]$. To disentangle this total effect into a direct and indirect (through M) causal channel, first note that the potential outcome can be rewritten as a function of both the treatment and the intermediate variable M: $Y(d) = Y(d, M(d))$. It follows that the (average) direct effect is identified by

$$\theta(d) = E[Y(1, M(d)) - Y(0, M(d))], \quad d \in \{0, 1\} \tag{1}$$

i.e. by exogenously varying the treatment but keeping the mediator fixed at its potential value for $D = d$. Equivalently, the (average) indirect effects is defined as

$$\delta(d) = E[Y(d, M(1)) - Y(d, M(0))], \quad d \in \{0, 1\} \tag{2}$$

i.e. by exogenously shifting the mediator to its potential values under treatment and non-treatment but keeping the treatment fixed at $D = d$.[4] Note that the ATE is the sum of the direct and indirect effects defined upon opposite treatment states:

[4] Pearl (2001) named these parameters the natural direct and indirect effects, whereas Robins (2003) referred to them as the pure direct and indirect effects and Flores and Flores-Lagunes (2009) as net and mechanism average treatment effects, respectively.

DOI: 10.1002/jae

$$\begin{aligned}\Delta &= E[Y(1,M(1)) - Y(0,M(0))] \\ &= E[Y(1,M(1)) - Y(0,M(1))] + E[Y(0,M(1)) - Y(0,M(0))] = \theta(1) + \delta(0) \\ &= E[Y(1,M(0)) - Y(0,M(0))] + E[Y(1,M(1)) - Y(1,M(0))] = \theta(0) + \delta(1)\end{aligned} \tag{3}$$

which follows from adding and subtracting $E[Y(0,M(1))]$ after the second and $E[Y(1,M(0))]$ after the fourth equality. The notation $\theta(1), \theta(0)$ and $\delta(1), \delta(0)$ highlights the possibility of effect heterogeneity w.r.t. the treatment state, i.e. the presence of interaction effects between the treatment and the mediator. However, it is obvious that these effects cannot be identified without further assumptions, as either $Y(1,M(1))$ or $Y(0,M(0))$ is observed for any unit, whereas $Y(1,M(0))$ and $Y(0,M(1))$ are never observed. Therefore, identification of direct and indirect effects hinges on the existence of exogenous variation in the treatment and the mediator.

2.2. Identification Given Observed Confounders not Affected by the Treatment

We now introduce our identifying assumptions, maintaining an i.i.d. framework throughout the paper. We start with the framework of conditional mediator exogeneity given the treatment and observed covariates (denoted by X) which are themselves *not* a function of D, with the leading case being pre-treatment covariates (evaluated prior to treatment assignment). Figure 1 provides a graphical illustration using a directed acyclic graph, where each arrow represents a causal path. Below, we will consider another set of restrictions assuming conditional mediator exogeneity given the treatment and covariates that are (at least partially) themselves a function of D, and thus post-treatment variables, which makes identification more difficult.

Our first assumption requires the treatment to be conditionally independent (given X) of any potential post-treatment variable, i.e. the potential mediator states and the potential outcomes. This is referred to as conditional independence, selection on observables, or exogeneity in the treatment evaluation literature (see, for instance, Imbens, 2004).

Assumption 1. (Conditional independence of the treatment)
$\{Y(d', m), M(d)\} \perp D|X$ for all $d', d \in \{0,1\}$ and m in the support of M.

Assumption 1 implies that there are no unobserved confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand, conditional on the covariates X. In observational studies the plausibility of this assumption, which has been criticized, among others, by Heckman and Navarro-Lozano (2004), critically hinges on the richness of the data. In experiments, it is satisfied if the treatment is either randomized within strata defined on X or randomized unconditionally, i.e. independent of X. (In the latter case, even the stronger condition $\{Y(d', m), M(d), X\} \perp D$ holds.)

The second assumption imposes conditional independence of the mediator given the treatment and the covariates along with a common support restriction on the conditional treatment probability.

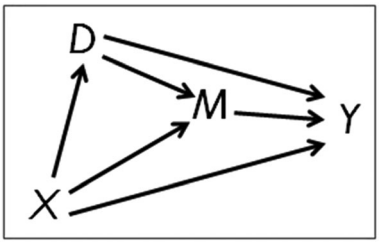

Figure 1. Causal paths under conditional exogeneity of the mediator

DOI: 10.1002/jae

Assumption 2. (Conditional independence of the mediator)

a. $Y(d', m) \perp M|D = d, X = x$ for all $d', d \in \{0,1\}$ and m, x in the support of M, X;
b. $\Pr(D = d|M = m, X = x) > 0$ for all $d \in \{0,1\}$ and m, x in the support of M, X.

Assumption 2 (a) states that, conditional on D and X, the effect of the mediator on the outcome is unconfounded. This implies that there exist no unobserved confounders jointly causing the mediator and the outcome given the treatment and the covariates. Assumption 2(a) is, for instance, violated if unobserved pre-treatment variables affect both M and Y directly, i.e. not only through D and X. Neither does it hold if unobserved post-treatment variables influence M and Y and are not fully determined by X and/or D. Therefore, 2(a) is a very strong restriction and only appears realistic if detailed information on potential confounders of themediator is available in the data (even in experiments with random treatment assignment where Assumption 1 is naturally satisfied). Furthermore, the described issue of post-treatment confounding must be plausibly ruled out. Note that, alternatively to 2(a), identification might also be based on an instrument for the mediator as, for instance, discussed in Imai *et al*. (2011).

Assumption 2(b) is a common support restriction requiring that the conditional probability to be treated given M, X, henceforth referred to as propensity score, is larger than zero in either treatment state. It follows that $\Pr(D = d|X = x) > 0$ must hold, too. Note that, by Bayes' theorem, Assumption 2(b) also implies that $\Pr(M = m|D = d, X = x) > 0$ (or, in the case of M being continuous, that the conditional density of M given D, X is larger than zero: $f_{M|D,X}(m,d,x) > 0$). That is, conditional on X, the mediator state must not be a deterministic function of the treatment, otherwise identification is infeasible due to the lack of comparable units in terms of the mediator across treatment states. Assumptions 1 and 2 correspond to the sequential ignorability assumption of Imai *et al*. (2010) (their Assumption 1) and Tchetgen Tchetgen and Shpitser (2011b), among others (see also the closely related Theorem 2 of Pearl, 2001). Also, Flores and Flores-Lagunes (2009) use a similar set of restrictions in their Assumptions 2, 5 and 6, but in addition impose a functional form assumption (their Assumption 3) which is not necessary for nonparametric identification.

Assumptions 1 and 2 allow us to identify $E[Y(d, M(d))]$ and $E[Y(d, M(1-d))]$. Starting with the former, note that

$$\begin{aligned} E[Y(d, M(d))] &= E[E[Y(d, M(d))|X = x]] = E[E[Y|D = d, X = x]] \\ &= E\left[E\left[\frac{Y \cdot I\{D = d\}}{\Pr(D = d|X)}|X = x\right]\right] = E\left[\frac{Y \cdot I\{D = d\}}{\Pr(D = d|X)}\right] \end{aligned} \tag{4}$$

where the first equality follows from the law of iterated expectations, the second from Assumption 1, the third from basic probability theory and the last from the law of iterated expectations (see also Hirano *et al*., 2003). Concerning the latter:

$$\begin{aligned} & E[Y(d, M(1-d))] \\ &= \int\int E[Y(d,m)|M(1-d) = m, X = x] dF_{M(1-d)|X=x}(m) \mathrm{d}F_X(x) \\ &= \int\int E[Y(d,m)|D = d, M = m, X = x] dF_{M|D=1-d,X=x}(m) \mathrm{d}F_X(x) \\ &= \int\int E[Y|D = d, M = m, X = x] \cdot \frac{\Pr(D = 1-d|M,X)}{\Pr(D = 1-d|X)} \mathrm{d}F_{M|X=x}(m) \mathrm{d}F_X(x) \\ &= E\left[E\left[E\left[\frac{Y \cdot I\{D = d\}}{\Pr(D = d|M,X)}|M = m, X = x\right] \cdot \frac{\Pr(D = 1-d|M,X)}{\Pr(D = 1-d|X)}|X = x\right]\right] \\ &= E\left[\frac{Y \cdot I\{D = d\}}{\Pr(D = d|M,X)} \cdot \frac{\Pr(D = 1-d|M,X)}{\Pr(D = 1-d|X)}\right] \end{aligned} \tag{5}$$

DOI: 10.1002/jae

The first equality follows from the law of iterated expectations and from replacing the outer expectations by integrals, the second from Assumptions 1 and 2, the third from Bayes' theorem, the fourth from basic probability theory and from replacing the integrals by expectations and the last from the law of iterated expectations. Therefore, $\theta(d)$ and $\delta(d)$ are identified by either subtracting equation (5) from equation (4) or vice versa, depending on whether d is one or zero. It follows by simple algebra that the direct and indirect effects are obtained from Propositions 1 and 2.[5]

Proposition 1. Under Assumptions 1 and 2, the average direct effect is identified by

$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d|X)}\right] \tag{6}$$

Proposition 1 implies that by propensity score-based weighting the distributions of both M and X are balanced between treatment and control groups such that the direct effect is identified. In particular, the distribution of the mediator in both groups corresponds to that of *M(d)* in the total population. Concerning the indirect effect, note that by equation (3) it corresponds to the difference between the average and the direct effect defined on the opposite treatment state: $\delta(d) = \Delta - \theta(1-d)$. Proposition 2 provides the representation of the indirect effect based on IPW, which is numerically identical to this difference.

Proposition 2. Under Assumptions 1 and 2, the average indirect effect is identified by

$$\delta(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,X)} \cdot \left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right] \tag{7}$$

From a practitioner's perspective, a nice feature of these identification results is that they are straightforward to implement. They only involve the (possibly parametric or nonparametric) estimation of two binary choice models for the propensity scores which are then plugged into the (normalized) sample analogs of Propositions 1 and 2, as outlined in Section 3. No parametric restrictions are imposed on the models of the outcome and the mediator such that arbitrary nonlinearities are allowed for. In contrast, the standard approach in the literature consists in estimating the ingredients of the following alternative representations of the parameters of interest; see, for instance, equations (8) and (26) in Pearl (2001) and Theorem 1 in Imai *et al*. (2010):

[5] Propositions 1 and 2 can also be derived by starting from the mediation formulae (see Pearl, 2001) provided in equation (8), which identify the direct and indirect effects under Assumptions 1 and 2. For example, considering the direct effect, note that

$$\begin{aligned}
\theta(d) &= \int\int \{E[Y|D=1,M=m,X=x] - E[Y|D=0,M=m,X=x]\} \mathrm{d}F_{M|D=d,X=x}(m)\mathrm{d}F_X(x) \\
&= \int\int \{E[Y|D=1,M=m,X=x] - E[Y|D=0,M=m,X=x]\} \frac{\Pr(D=d|M,X)}{\Pr(D=d|X)} \mathrm{d}F_{M|X=x}(m)\mathrm{d}F_X(x) \\
&= E\left[E\left[E\left[\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right] \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d|X)} | X=x\right]\right] \\
&= E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,X)}\right) \cdot \frac{\Pr(D=d|M,X)}{\Pr(D=d|X)}\right]
\end{aligned}$$

DOI: 10.1002/jae

$$\theta(d) = \int\int \{E[Y|D=1, M=m, X=x] - E[Y|D=0, M=m, X=x]\} \mathrm{d}F_{M|D=d,X=x}(m) \mathrm{d}F_X(x) \quad (8)$$

$$\delta(d) = \int\int E[Y|D=d, M=m, X=x] \left\{\mathrm{d}F_{M|D=1,X=x}(m) - \mathrm{d}F_{M|D=0,X=x}(m)\right\} \mathrm{d}F_X(x) \quad (9)$$

This requires estimators for the conditional mean of Y given D, M, X and the conditional density of M given D, X. In the literature, parametric methods have been most commonly used (see, for instance, Pearl, 2011; VanderWeele, 2009).[6] They, however, appear unattractive due to their severe functional form restrictions and the potentially difficult interpretability of direct and indirect effects under nonlinear modeling (e.g. when both the outcome and the mediator are binary). Nonparametric estimation, as recently proposed in Imai *et al*. (2010), avoids these shortcomings but might be cumbersome in empirical applications if X is high-dimensional and/or M is continuous. In contrast, estimation based on Propositions 1 and 2 is less prone to such issues as it just relies on two propensity score models. Our IPW-based results are also more general than the regression approach of Flores and Flores-Lagunes (2009). The latter does not require the estimation of the conditional density of the mediator, but imposes a functional form restriction (their Assumption 3) on the expected potential outcomes across potential mediator states (for the treatment fixed) which we need not invoke here.[7]

However, it is important to note that IPW also has its drawbacks: if the common support Assumption 2(b) is close to being violated, estimation may be unstable and the variance may explode (see Frölich (2004) and Khan and Tamer (2010), among others). Furthermore, IPW is less robust to propensity score misspecification than other classes of estimators, as documented, for instance, in Kang and Schafer (2007) and Waernbaum (2012). Therefore, matching on the propensity score (see, for instance, Rosenbaum and Rubin, 1983), which remains consistent under particular forms of propensity score misspecification, might represent a viable alternative to IPW.

2.3.. Identification Given Observed Confounders Affected by the Treatment

It appears unlikely in many applications that conditioning on pre-treatment covariates is sufficient to control for mediator endogeneity, given that the mediator is itself a post-treatment variable. Equivalent to the treatment evaluation literature, where potential confounders of the treatment are measured at or shortly before the treatment, potential confounders of the mediator should be controlled for just before the selection into the mediator takes place. Then, however, it appears likely that at least some of these covariates are a function of the treatment, too, implying that they are themselves mediators that affect the mediator of interest. Therefore, Robins (2003) suspects that the set-up relying on Assumptions 1 and 2 is of limited practical relevance. This most likely also applies to our application presented in Section 5, where we are interested in the effect of the Job Corps program on health. The mediator is employment and, clearly, some potential confounders affecting both employability and health (such as the labor market state shortly prior to employment) are most likely a function of the treatment.

As in Robins (2003) and Imai and Yamamoto (2013), we therefore also consider a framework in which D is permitted to have an effect on post-treatment confounders of the mediator, which we denote by W. In this case, mediation analysis becomes more complicated and requires us to introduce additional notation by rewriting the potential mediator and potential outcome also as functions of

[6] Furthermore, Tchetgen Tchetgen and Shpitser (2011a) and Zheng and van der Laan (2012), among others, provide multiply robust parametric estimators based on conditional mean estimation of the outcome and conditional density estimation of both the treatment and the mediator.

[7] Only if the mediator is exogenous conditional on post-treatment confounders which are themselves influenced by the treatment, do we have to rely on a similar assumption, as outlined below.

DOI: 10.1002/jae

W: $M(d) = M(d, W(d))$ and $Y(d, M(d)) = Y(d, M(d, W(d)), W(d))$, where $W(d)$ is the vector of potential values of W for $D = d$. Then, the total indirect effect is defined as

$$\delta^t(d) = E[Y(d, M(1, W(1)), W(d)) - Y(d, M(0, W(0)), W(d))] \tag{10}$$

We refer to $\delta^t(d)$ as the total indirect effect because it comprises all effects via M which either come from D directly or 'take a devious route' through W. That is, this parameter accounts for the fact that M is affected by D both directly and indirectly through a change in W. In contrast, the partial indirect effect only identifies the effect through M directly coming from D, but not going through W:

$$\delta^p(d) = E[Y(d, M(1, W(d)), W(d)) - Y(d, M(0, W(d)), W(d))] \tag{11}$$

That is, $\delta^p(d)$ is the *ceteris paribus* indirect effect via the mediator when holding W constant at the level implied by d such that any channel through the post-treatment covariates is shut down. This is what regressing Y on (1,D,M,W) and multiplying the coefficient on M with the first stage effect of D on M obtains as 'indirect effect' in the limit, given that no further confounders are present. Obviously, this effect neglects any correlations between M and W. We therefore argue that the total indirect effect is the more interesting parameter,[8] but nevertheless discuss the identification of both parameters. However, it will be shown further below that $\delta^p(d)$ is more easily identified than $\delta^t(d)$. As a further remark, note that $\delta^t(d)$ corresponds to $\delta(d)$ (which also refers to the total of indirect effects) in Section 2.2, where we, however, did not distinguish between total and partial effects due to the absence of post-treatment confounders.

The direct effect is defined as

$$\theta(d) = E[Y(1, M(d, W(d)), W(d)) - Y(0, M(d, W(d)), W(d))] \tag{12}$$

i.e. it corresponds to the change in the mean potential outcome due to an exogenous change in the treatment, while keeping the mediator and the post-treatment covariates fixed. Note that this definition differs from Imai and Yamamoto (2013), who consider the difference between the ATE and the total indirect effect to be the 'direct' effect: $E[Y(d, M(d, W(d)), W(d)) - Y(1-d, M(d, W(d)), W(1-d))]$. However, this includes changes in the mean potential outcome which are due to a change in W which is not mediated by M. Here, we define the direct effect in a narrower sense that also excludes (inherently indirect) channels via W. For this reason, $\theta(d)$ and $\delta^t(d)$ or $\delta^p(d)$, respectively, do not add up to the ATE, as either $E[Y(d, M(d, W(1)), W(1)) - Y(d, M(d, W(0)), W(0))]$ or $E[Y(d, M(d, W(d)), W(1)) - Y(d, M(d, W(d)), W(0))]$ are not accounted for, respectively.

The directed acyclic graph in Figure 2 displays a set-up where the treatment affects the observed confounders W of the mediator. Because we need to condition on W, identification requires ruling out unobserved confounders that jointly cause W on the one hand and M and/or Y on the other hand. Furthermore, just as in Section 2.2, we also have to control for pre-treatment covariates that jointly affect the treatment and the mediator/outcome. A further identification issue arises if all or some pre-treatment covariates also have an impact on the post-treatment covariates W. Then, conditioning on the latter changes the distribution of pre-treatment covariates even if they were initially balanced across treatment states as in randomized experiments. For this reason, we, in addition to the confounders of the treatment, also need to control for all those pre-treatment covariates that are jointly related to W on the one hand and directly to the outcome and/or the mediator on the other. Otherwise, conditioning on W would introduce a correlation between D and those pre-treatment covariates predicting both W and M or Y and, thus, treatment endogeneity. We denote by X the vector of all pre-treatment covariates

[8] Also, Imai and Yamamoto (2013) focus on $\delta^t(d)$ as indirect effect and do not consider $\delta^p(d)$ at all.

DOI: 10.1002/jae

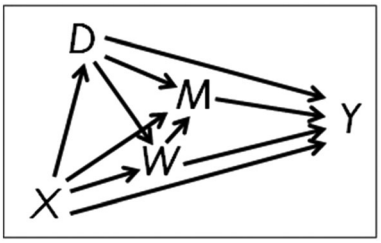

Figure 2. Causal paths with pre-treatment covariates (X) and post-treatment covariates (W)

that may cause one or several of the forms of confounding just discussed and replace Assumptions 1 and 2 by Assumptions 3 and 4 to incorporate post-treatment confounders of the mediator.

Assumption 3. (Conditional independence of the treatment)

a. $\{Y(d'',m,w'), M(d',w), W(d)\} \perp D|X=x$ for all $d'', d', d \in \{0,1\}$ and m, w', w, x in the support of M, W, X;
b. $\{Y(d'',m,w''), M(d',w')\} \perp D|W=w, X=x$ for all $d', d \in \{0,1\}$ and m, w'', w', w, x in the support of M, W, X.

Similar to Assumption 1, Assumption 3(a) says that D is conditionally independent of any potential post-treatment variables, namely the potential outcomes, the potential mediator states and the potential values of the post-treatment confounders of the mediator. Assumption 3(b) is new and states that the conditional independence of the treatment and the potential outcomes or mediator states must also hold when jointly conditioning on W and X, implying that all pre-treatment covariates affecting both W and M or Y are included in X.

Assumption 4. (Conditional independence of the mediator)

a. $Y(d'',m,w') \perp M|D=d, W=w, X=x$ for all $d'', d', d \in \{0,1\}$ and m, w', w, x in the support of M, W, X;
b. $Pr(D=d|M=m, W=w, X=x) > 0$ for all $d \in \{0,1\}$ and m, w, x in the support of M, W, X.

Assumption 4(a) is related to, but somewhat weaker than Assumption 2(a), as conditional independence is now only required to hold given pre- and post-treatment covariates. This, for instance, rules out that unobserved pre-treatment variables affect both M and Y directly, i.e. not only through D, X and/or W. Furthermore, it requires that any unobserved post-treatment variables which are not a deterministic function of D and/or X must not jointly cause M and Y conditional on W. Assumption 4(b) says that the common support restriction on the treatment propensity must hold when conditioning on M and both W and X.

When comparing our assumptions to others made in the literature, it turns out that Assumptions 3 and 4(a) are similar to FRCISTG (fully randomized causally interpretable structural tree graph) in Robins (2003) (see also Robins, 1986) and Assumption 2 in Imai and Yamamoto (2013), with one important difference: Robins (2003) and Imai and Yamamoto (2013) merely assume that $Y(d, m, W(d))$ and $M(d)$ defined on the same treatment are conditionally independent. In contrast, Assumption 4(a) imposes conditional independence of $Y(d'', m, W(d'))$ and $M(d)$ for possibly distinct d'', d', d (i.e. also for potential outcomes and potential mediator states defined on opposite treatments).[9] This

[9] Robins and Richardson (2010) present a DGP in their Appendix B, where indeed FRCISTG holds, while our stronger restriction is not satisfied. However, Robins (2003) argues that it seems hard to construct realistic scenarios where one set of assumptions holds while the other one does not.

DOI: 10.1002/jae

stronger conditional independence allows for nonparametric identification of the direct and partial indirect effects without further restrictions (other than Assumptions 3 and 4), but not of the total indirect effect, as discussed below. In contrast, FRCISTG does not identify any of the parameters without further functional form restrictions (as imposed in Robins, 2003, and Imai and Yamamoto, 2013).

Propositions 3 and 4 concern the identification of the direct effect, which hinges on the identifiability of $E[Y(1-d, M(d, W(d)), W(d))]$, and of the partial indirect effect, respectively, which requires $E[Y(d, M(1-d, W(d)), W(d))]$. The respective proofs, which are similar to equation (5), are provided in Appendix A.2 (supporting information).

Proposition 3. Under Assumptions 3 and 4, the average direct effect is identified by

$$\theta(d) = E\left[\left(\frac{Y \cdot D}{\Pr(D=1|M,W,X)} - \frac{Y \cdot (1-D)}{1-\Pr(D=1|M,W,X)}\right) \cdot \frac{\Pr(D=d|M,W,X)}{\Pr(D=d|X)}\right] \tag{13}$$

Proposition 4. Under Assumptions 3 and 4, the average partial indirect effect is identified by

$$\delta^p(d) = E\left[\frac{Y \cdot I\{D=d\}}{\Pr(D=d|M,W,X)} \cdot \frac{\Pr(D=d|W,X)}{\Pr(D=d|X)} \cdot \left(\frac{\Pr(D=1|M,W,X)}{\Pr(D=1|W,X)} - \frac{1-\Pr(D=1|M,W,X)}{1-\Pr(D=1|W,X)}\right)\right] \tag{14}$$

The identification of the total indirect effect requires identifying $E[Y(d, M(1-d, W(1-d)), W(d))]$. Unfortunately, this is not feasible without further assumptions; see the proof in Avin *et al.* (2005) and the results of Albert and Nelson (2011), who consider a sequential conditional independence assumption that is similar to ours. The reason is that conditional on X (when the treatment is as good as random), the identification of $E[Y(d, M(1-d, W(1-d)), W(d))]$ requires exogenously adjusting the distribution of M given $D=d$ to that of M given $D=1-d$, while at the same time keeping the distribution of W fixed (given $D=d$). Obviously, this is impossible if W and M are not independent conditional on X.

However, Robins (2003) shows under FRCISTG that the total indirect effect is identified under an additional restriction, namely the absence of interaction effects between D and M. Formally, his assumption implies that the unit-level treatment effect (for any unit i) for the mediator fixed is constant across different values of the mediator:

$$Y_i(1, m, X_i(1)) - Y_i(0, m, X_i(0)) = Y_i(1, m', X_i(1)) - Y_i(0, m', X_i(0)) = B_i$$

where B_i is a unit-level constant. Unfortunately, this assumption appears unattractive in empirical applications (see, for instance, the discussion in Imai *et al.*, 2012, Section 3.1) and restricts the usefulness of nonparametric identification advocated in recent work. However, Imai and Yamamoto (2013) demonstrate that under FRCISTG the assumption of no interaction effect can be relaxed to assuming a homogeneous interaction effect:

$$Y_i(1, m, X_i(1)) - Y_i(0, m, X_i(0)) = B_i + Cm$$

where C is constant for any m, i.e. the interaction between the treatment and the mediator varies homogeneously for all observations.

DOI: 10.1002/jae

Here, we propose a functional form restriction on the relation of the mean potential outcome and the mediator across treatment states that is comparable to Assumption 3 in Flores and Flores-Lagunes (2009) (who, however, use it in the set-up where D does not affect W).

Assumption 5. (Functional form of mean potential outcome–mediator relation)

a. $E[Y(d,m,w)|D=d, M(d, W(d))=m, W(d)=w, X=x]=\mu(d,m,w,x)$ for all $d\in\{0,1\}$ and m, w, x in the support of M, W, X, with the function $\mu(D,M,W,X)$ being linear in M;
b. $E[Y(d,m,w)|D=d, M(d, W(d))=m, W(d)=w, X=x]=E[Y(d,m,w)|D=d, M(1-d, W(1-d))=m, W(d)=w, X=x]$ for all $d\in\{0,1\}$ and m, w, x in the support of M, W, X.

By Assumption 5(a) the conditional mean potential outcome under $D=d$ can be characterized by the regression model μ, which is restricted to be a linear function of the mediator. Assumption 5(b) states that one can use this very same regression model to predict $E[Y(d,m,x)|M(1-d, W(1-d))=m, W(d)=w, X=x]$. That is, the functional form of $\mu(d, M(d, W(d)), W(d), X)$ is assumed to be identical to $\mu(d, M(1-d, W(1-d)), W(d), X)$ whenever $M(d, W(d))=M(1-d, W(1-d))$. This restriction implies that the mean interaction effect of the treatment and the mediator is the same for any $M(1, W(1))=M(0, W(0))=m$ conditional on $W(d)$ and X, otherwise the regression functions necessarily differ. Together with Assumptions 3 and 4, Assumption 5 permits the identification of $E[Y(d, M(1-d, W(1-d)), W(d))]$, even though $M(1-d, W(1-d))$ is not known for units with $D=d$. The trick is to use $E[M(1-d, W(1-d))]$, which can be identified in the population with $D=1-d$, in our regression model: $\mu(d, E[M(1-d, W(1-d))], W(d), x)$. By the linearity assumption 5(b) and the law of iterated expectations it follows that $E[\mu(d, E[M(1-d, W(1-d))], w, x)|D=d]=E[\mu(d, M(1-d, W(1-d)), w, x)|D=d]$, which allows identifying $E[Y(d, M(1-d, W(1-d)), W(d))]$. However, it has to be stressed that Assumption 5 is far from being innocuous. Firstly, it requires a correctly specified regression model that allows making predictions across mediator states. Secondly, the linearity assumption rules out more general relations between the outcome and the mediator.

Given that these considerably stronger functional form assumptions are satisfied, $E[Y(d, M(1-d, W(1-d)), W(d))]$, which is required for identifying the total indirect effect, is obtained from the following result:

$$
\begin{aligned}
& E[Y(d, M(1-d, W(1-d)), W(d))] \\
&= \int E[E[Y(d, M(1-d, W(1-d)), w)|W(d)=w, X=x]|X=x]dF_X(x) \\
&= \int E[E[Y(d, M(1-d, W(1-d)), w)|D=d, W=w, X=x]|D=d, X=x]\mathrm{d}F_X(x) \\
&= \int E[E[Y(d, E[M(1-d, W(1-d))], w)|D=d, W=w, X=x]|D=d, X=x]\mathrm{d}F_X(x) \\
&= \int E[E[Y(d, E[M(1-d, W(1-d))], w)|D=d, M=E[M(1-d, W(1-d))], W \\
&= w, X=x]|D=d, X=x]\mathrm{d}F_X(x) \\
&= \int E[\mu(d, E[M(1-d, W(1-d))], W, x)|D=d, X=x]\mathrm{d}F_X(x) \\
&= \int E\left[\frac{\mu\left(d, E\left[E\left[\frac{M\cdot I\{D=1-d\}}{\Pr(D=1-d|X)}|X=x\right]\right], W, x\right)\cdot I\{D=d\}}{\Pr(D=d|X)}|X=x\right]\mathrm{d}F_X(x) \\
&= E\left[\frac{\mu\left(d, E\left[\frac{M\cdot I\{D=1-d\}}{\Pr(D=1-d|X)}\right], W, X\right)\cdot I\{D=d\}}{\Pr(D=d|X)}\right]
\end{aligned}
\tag{15}
$$

DOI: 10.1002/jae

The first equality follows from the law of iterated expectations and from replacing the outer expectation by an integral, the second from Assumption 3, the third from Assumption 5, the fourth from Assumption 4, the fifth from Assumption 5, the sixth from Assumption 3 and from basic probability theory, and the last from replacing the integral by an expectation and from the law of iterated expectations.

Proposition 5. Under Assumptions 3, 4 and 5, the average total indirect effect is identified by

$$\delta^t(1) = E\left[\frac{\left\{Y - \mu\left(1, E\left[\frac{M\cdot(1-D)}{1-\Pr(D=1|X)}\right], W, X\right)\right\}\cdot D}{\Pr(D=1|X)}\right], \tag{16}$$

$$\delta^t(0) = E\left[\frac{\left\{\mu\left(0, E\left[\frac{M\cdot D}{\Pr(D=1|X)}\right], W, X\right) - Y\right\}\cdot(1-D)}{1-\Pr(D=1|X)}\right]$$

It is interesting to compare this result to Albert and Nelson (2011), who use a similarly strong sequential ignorability assumption and assume parametric models for D, W, M and Y, which are estimated by maximum likelihood methods. However, they do not identify $E[Y(d, M(1-d, W(1-d)), W(d))]$ and $\delta^t(d)$ without further assumptions, because the linearity restriction (5a) is not imposed therein. Therefore, recovering the total indirect effect comes at thecost of ruling out models that are nonlinear in the mediator, which are permitted in Albert and Nelson (2011).

3. ESTIMATION

For estimation in Sections 4 and 5, we use normalized versions of the sample analogs of the IPW-based identification results in Propositions 1–4, such that the weights of the observations in either treatment state add up to unity, as advocated in Imbens (2004) and Busso *et al*. (2009b).[10] For example, the normalized estimators of the direct effects under treatment and non-treatment are given by

$$\hat{\theta}(1) = \frac{\sum Y_i\cdot D_i/\hat{p}(X_i)}{\sum D_i/\hat{p}(X_i)} - \frac{\sum Y_i\cdot(1-D_i)\cdot\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\cdot\hat{p}(X_i)]}{\sum(1-D_i)\cdot\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\cdot\hat{p}(X_i)]},$$

$$\hat{\theta}(0) = \frac{\sum Y_i\cdot D_i\cdot(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)\cdot(1-\hat{p}(X_i))]}{\sum D_i\cdot(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)\cdot(1-\hat{p}(X_i))]} - \frac{\sum Y_i\cdot(1-D_i)/(1-\hat{p}(X_i))}{\sum(1-D_i)/(1-\hat{p}(X_i))}$$

where i is the index of the observations in the i.i.d. sample and $\hat{p}(M_i, X_i)$, $\hat{p}(X_i)$ denote the respective estimates of the propensity scores $\Pr(D=1|M_i, X_i)$, $\Pr(D=1|X_i)$, which we estimate by probit specifications. Our semiparametric IPW estimators (into which the propensity scores enter parametrically) can be expressed as sequential GMM estimators where propensity score estimation represents the first step and effect estimation the second step (see Newey, 1984). It follows from his results that our methods are $\sqrt{n}$ -consistent under standard regularity conditions. Furthermore, IPW estimators are sufficiently smooth for the bootstrap being consistent for inference. We therefore estimate the standard errors in the application in Section 5 by 1999 bootstrap draws.

[10] We do not use any propensity score trimming as considered in Busso *et al*. (2009a), Crump *et al*. (2009) and Huber *et al*. (2013), because propensity scores close to the boundaries 0 and 1 occur neither in our simulation study in Section 4 nor in our application in Section 5.

DOI: 10.1002/jae

Concerning the estimation of the total indirect effects based on Proposition 5, μ is specified as a linear function of the covariates and the mediator within each treatment state, such that two linear models (for $D=1,0$) have to be estimated. That is, $\mu(d,M,X,W) = \beta_0^d + M\beta_M^d + X\beta_X^d + W\beta_W^d$, where $a_0^d, \beta_M^d, \beta_X^d, \beta_W^d$ represent the constant and the slope coefficients on M,X,W for $D=d$. Note that estimating two separate models for each treatment state allows for interactions of D and each of M,X, W. For inference, we again use the bootstrap in the application.

4. SIMULATIONS

This section presents a simulation study that provides some intuition for the identification results and the issues related to imposing the wrong set of assumptions. For ease of exposition, we consider a data-generating process (DGP) which is based on linear equations:

$$Y = 0.5D + M + \beta DM + V + 0.25U + \varepsilon_1 \tag{17}$$

$$M = 0.5D + 0.5V + 0.25U + \varepsilon_2 \tag{18}$$

$$V = \gamma D + 0.25U + \varepsilon_3 \tag{19}$$

$$D = I\{0.25U + \varepsilon_4 > 0\}, \text{with } U, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4 \sim N(0,1), \text{independently of each other} \tag{20}$$

Equation (17) is the outcome equation, in which the observed Y is a function of the observed variables D, M, V, U and an unobserved term ε_1. β gauges the interaction effect between D and M such that $\beta=0$ satisfies the assumption of no interaction discussed in Robins (2003). By equation (18), the mediator is a function of D, V, U and the unobservable ε_2.[11] The parameter γ in equation (19) determines whether V is caused by D and which set of assumptions is valid. If $\gamma=0$, Assumptions 1 and 2 hold if X therein is defined as $\{V,U\}$. Otherwise, identification has to be based on Assumptions 3–5, with W therein corresponding to V and X corresponding to U. By equation (20), the treatment is conditionally independent given U with an unconditional treatment probability of 0.5. In the simulations, we set γ to 0 and 0.2 and β to 0 and 0.5. Table I provides the true direct and indirect effects for the various scenarios, which are explained in more detail in Appendix A.3 (supporting information).

We run 5000 Monte Carlo simulations with 2000 observations and estimate the models by the normalized sample analogs of either Propositions 1 and 2 (for $\gamma=0$) or Propositions 3–5 (for $\gamma=0.2$), respectively. In addition, we also consider ordinary least squares (OLS) regression, which estimates the direct effect as the coefficient on D in a regression of Y on (1,D,M,V,U). The indirect effect corresponds to the coefficient on M in the latter regression multiplied by the coefficient on D in a regression of M on (1,D,V,U). This approach omits interactions between D and M so that direct and indirect effects are implicitly assumed to be homogeneous w.r.t. the treatment.[12] For this reason we also estimate a linear model that includes the treatment mediator interaction as modeled in equation (17) for $\beta \neq 0$. See, for instance, Imai *et al.* (2010) for a definition of the direct and indirect effects in terms ofregression coefficients in this somewhat more general framework. Finally, we also include a naive OLS estimator where the *D-M* interaction and V, U are omitted so that the confounders of the mediator/treatment are not controlled for.

[11] Note that the potential mediator states are $M(1) = 0.5 + 0.5V + 0.25U + \varepsilon_2$ and $M(0) = 0.5V + 0.25U + \varepsilon_2$, the potential outcomes are $Y(1, M(d)) = 0.5 + (1+\beta)M(d) + V + 0.25U + \varepsilon_1$ and $Y(0, M(d)) = M(d) + V + 0.25U + \varepsilon_1$ for $d \in \{1,0\}$.

[12] This issue is often encountered in empirical work; see Section 4.1 of Imai *et al.* (2011) for a revision and discussion of the shortcomings of the standard linear framework.

DOI: 10.1002/jae

Table I. True direct and indirect effects for various scenarios

	$\gamma=0$				$\gamma=0.2$			
	$\beta=0$		$\beta=0.5$		$\beta=0$		$\beta=0.5$	
Effect	$d=1$	$d=0$	$d=1$	$d=0$	$d=1$	$d=0$	$d=1$	$d=0$
$\theta(d)$	0.5	0.5	0.75	0.5	0.5	0.5	0.8	0.5
$\delta(d)$	0.5	0.5	0.75	0.5	—	—	—	—
$\delta^t(d)$	—	—	—	—	0.6	0.6	0.9	0.6
$\delta^p(d)$	—	—	—	—	0.5	0.5	0.75	0.5

Table II presents the bias, variance and mean squared error (MSE) of the various estimators for $\gamma=0$. The IPW estimators based on Propositions 1 and 2 ($\hat{\theta}_{\mathrm{IPW}}(d)$, $\hat{\delta}_{\mathrm{IPW}}(d)$) are almost unbiased and their MSEs are moderate in any scenario. The same applies to OLS regression, including the *D–M* interaction term ($\hat{\theta}_{\mathrm{OLS.ia}}(d)$, $\hat{\delta}_{\mathrm{OLS.ia}}(d)$), which is somewhat more efficient because it is fully parametric. For $\beta=0$, also the OLS estimators without interaction ($\hat{\theta}_{\mathrm{OLS}}(d)$, $\hat{\delta}_{\mathrm{OLS}}(d)$) perform decently. However, they are biased if $\beta=0.5$ due to omitting the interaction term. The naive estimators ($\hat{\theta}_{\mathrm{naive}}(d)$, $\hat{\delta}_{\mathrm{naive}}(d)$) are biased in general, as they disregard confounding.

Table III provides the results for $\gamma=0.2$. Concerning the direct effects, we observe a similar pattern in terms of the relative performance across estimators as before. Taking a look at the total indirect effect reveals that the estimator based on Proposition 4, $\hat{\delta}^t(d)$, is almost unbiased. In contrast, when using IPW based on Proposition 2 ($\hat{\delta}_{\mathrm{IPW}}(d)$) for estimating $\delta^t(d)$, the bias is non-negligible. Also, OLS both with and without treatment-mediator interaction is generally biased under any β, because V is controlled for in the regression as if it was not affected by D. However, for $\beta=0$, $\hat{\delta}_{\mathrm{OLS}}(d)$ still consistently estimates the partial indirect effect $\delta^p(d)$, i.e. the *ceteris paribus* impact going through M for V, U fixed. $\hat{\delta}_{\mathrm{OLS.ia}}(d)$ does so even if $\beta=0.5$, because it accounts for the interaction effect. Also $\hat{\delta}^p_{\mathrm{IPW}}(d)$, the IPW estimator of the indirect effect based on Proposition 3, performs well when estimating the partial indirect effect, while $\hat{\delta}_{\mathrm{IPW}}(d)$ is again biased. The bias of the naive estimator is large for both the partial and total indirect effects.

In conclusion, the methods proposed in this paper perform well in the simulations if chosen according to the appropriate identifying assumptions. However, so does $\hat{\theta}_{\mathrm{OLS.ia}}(d)$, $\hat{\delta}_{\mathrm{OLS.ia}}(d)$ for $\gamma=0$. Therefore, the question arises whether there exist scenarios in which IPW outperforms a parametric estimator that is flexible enough to allow for treatment–mediator interactions. This may generally be the case if the parametric model is misspecified. To demonstrate this, we reconsider the DGP with $\gamma=0$ and $\beta=0.5$, but add the interaction DU to the right-hand side of equation (17) such that $\hat{\theta}_{\mathrm{OLS.ia}}(d)$, $\hat{\delta}_{\mathrm{OLS.ia}}(d)$ are based on a wrong outcome model. In a second set-up, we additionally include the interaction DV. The results in Table IV show that IPW is still consistent, while the bias of $\hat{\theta}_{\mathrm{OLS.ia}}(d)$, $\hat{\delta}_{\mathrm{OLS.ia}}(d)$ increases as the misspecification of the outcome becomes more severe. A more comprehensive investigation of the finite sample performance of IPW relative to alternative estimators is left to future research.

5. APPLICATION

We apply the estimators resulting from Propositions 1–5 to a welfare policy experiment with a binary treatment assignment (D) which was conducted in the mid 1990s to assess the publicly funded US Job

DOI: 10.1002/jae

Table II. Bias, variance and MSE of various estimators under Assumptions 1 and 2 ($\gamma = 0$)

	$\beta = 0$						$\beta = 0.5$					
	$d = 1$			$d = 0$			$d = 1$			$d = 0$		
Est.	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
$\hat{\theta}_{\mathrm{IPW}}(d)$	−0.002	0.004	0.004	−0.002	0.004	0.004	−0.002	0.004	0.004	−0.003	0.005	0.005
$\hat{\theta}_{\mathrm{OLS.ia}}(d)$	−0.001	0.003	0.003	−0.001	0.003	0.003	−0.001	0.004	0.004	−0.000	0.004	0.004
$\hat{\theta}_{\mathrm{OLS}}(d)$	−0.000	0.002	0.002	−0.000	0.002	0.002	−0.125	0.002	0.018	0.125	0.002	0.018
$\hat{\theta}_{\mathrm{naive}}(d)$	−0.124	0.004	0.020	−0.124	0.004	0.020	−0.249	0.004	0.066	0.001	0.004	0.004
$\hat{\delta}_{\mathrm{IPW}}(d)$	0.000	0.008	0.008	0.000	0.008	0.008	0.000	0.013	0.013	0.000	0.008	0.008
$\hat{\delta}_{\mathrm{OLS.ia}}(d)$	−0.000	0.003	0.003	−0.000	0.003	0.003	−0.001	0.006	0.006	−0.000	0.003	0.003
$\hat{\delta}_{\mathrm{OLS}}(d)$	−0.001	0.003	0.003	−0.001	0.003	0.003	−0.126	0.004	0.020	0.124	0.004	0.020
$\hat{\delta}_{\mathrm{naive}}(d)$	0.461	0.006	0.219	0.461	0.006	0.219	0.373	0.009	0.147	0.623	0.009	0.396

DOI: 10.1002/jae

Table III. Bias, variance and MSE of various estimators under Assumptions 3–5 ($\gamma = 0.2$)

	$\beta = 0$						$\beta = 0.5$					
	$d = 1$			$d = 0$			$d = 1$			$d = 0$		
Est.	Bias	var	MSE	Bias	var	MSE	Bias	var	MSE	Bias	var	MSE
$\hat{\theta}_{\mathrm{IPW}}(d)$	−0.004	0.004	0.004	−0.003	0.004	0.004	−0.004	0.004	0.004	−0.004	0.006	0.006
$\hat{\theta}_{\mathrm{OLS.ia}}(d)$	−0.001	0.003	0.003	−0.001	0.003	0.003	−0.026	0.004	0.004	0.025	0.004	0.004
$\hat{\theta}_{\mathrm{OLS}}(d)$	−0.001	0.002	0.002	−0.001	0.002	0.002	−0.150	0.002	0.025	0.150	0.002	0.025
$\hat{\theta}_{\mathrm{naive}}(d)$	0.026	0.004	0.005	0.026	0.004	0.005	−0.123	0.004	0.020	0.177	0.004	0.036
$\hat{\delta}(d)$	−0.001	0.003	0.003	−0.001	0.003	0.003	−0.001	0.006	0.006	−0.001	0.003	0.003
$\hat{\delta}_{\mathrm{IPW}}(d)$ for $\delta^t(d)$	−0.099	0.009	0.018	0.202	0.009	0.049	0.202	0.014	0.055	0.202	0.009	0.049
$\hat{\delta}_{\mathrm{OLS.ia}}(d)$ for $\delta^t(d)$	−0.401	0.003	0.163	−0.100	0.003	0.013	−0.151	0.006	0.029	−0.100	0.003	0.013
$\hat{\delta}_{\mathrm{OLS}}(d)$ for $\delta^t(d)$	−0.100	0.002	0.012	−0.100	0.002	0.012	−0.275	0.003	0.079	0.025	0.003	0.004
$\hat{\delta}_{\mathrm{naive}}(d)$ for $\delta^t(d)$	0.511	0.006	0.267	0.511	0.006	0.267	0.397	0.009	0.166	0.697	0.009	0.494
$\hat{\delta}_{\mathrm{IPW}}(d)$	0.002	0.003	0.003	0.002	0.003	0.003	0.002	0.007	0.007	0.002	0.003	0.003
$\hat{\delta}_{\mathrm{IPW}}(d)$ for $\delta^p(d)$	0.051	0.009	0.011	0.302	0.009	0.100	0.352	0.014	0.138	0.302	0.009	0.100
$\hat{\delta}_{\mathrm{OLS.ia}}(d)$ for $\delta^p(d)$	−0.000	0.003	0.003	−0.000	0.003	0.003	−0.001	0.006	0.006	−0.000	0.003	0.003
$\hat{\delta}_{\mathrm{OLS}}(d)$ for $\delta^p(d)$	−0.000	0.002	0.002	−0.000	0.002	0.002	−0.125	0.003	0.019	0.125	0.003	0.019
$\delta_{\mathrm{naive}}(d)$ for $\delta^p(d)$	0.611	0.006	0.379	0.611	0.006	0.379	0.547	0.009	0.308	0.797	0.009	0.644

DOI: 10.1002/jae

Table IV. Bias, variance and MSE of various estimators under β =0.5 and $\gamma = 0.2$

Est.	interaction DU added to eq. (17)						Interactions $DU + DV$ added to eq. (17)					
	$d=1$			$d=0$			$d=1$			$d=0$		
	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
$\hat{\theta}_{\mathrm{IPW}}(d)$	−0.002	0.005	0.005	−0.003	0.006	0.006	−0.002	0.006	0.006	−0.002	0.008	0.008
$\hat{\theta}_{\mathrm{OLS.ia}}(d)$	0.064	0.005	0.009	−0.065	0.005	0.009	0.171	0.007	0.036	−0.172	0.007	0.036
$\hat{\theta}_{\mathrm{OLS}}(d)$	−0.125	0.003	0.019	0.125	0.003	0.019	−0.125	0.004	0.020	0.125	0.004	0.020
$\hat{\theta}_{\mathrm{naive}}(d)$	−0.140	0.006	0.026	0.110	0.006	0.018	−0.050	0.006	0.008	0.200	0.006	0.046
$\hat{\delta}_{\mathrm{IPW}}(d)$	0.001	0.014	0.014	0.000	0.008	0.008	0.001	0.014	0.014	0.001	0.003	0.003
$\hat{\delta}_{\mathrm{OLS.ia}}(d)$	0.064	0.007	0.011	−0.065	0.003	0.007	0.170	0.009	0.039	−0.172	0.002	0.031
$\hat{\delta}_{\mathrm{OLS}}(d)$	−0.125	0.003	0.019	0.125	0.003	0.019	−0.125	0.003	0.019	0.125	0.003	0.019
$\hat{\delta}_{\mathrm{naive}}(d)$	0.456	0.010	0.218	0.706	0.010	0.509	0.319	0.008	0.110	0.569	0.008	0.332

DOI: 10.1002/jae

Corps program.[13] The program targets young individuals (aged 16–24 years) who legally reside in the USA and come from a low-income household. It provides participants with approximately 1200 hours of vocational training and education, housing and board over an average duration of 8 months. Participants also receive health education as well as health and dental care. Schochet *et al*. (2001, 2008) discuss in detail the experimental design[14] and the ATEs on a broad range of outcomes. Their findings suggest that Job Corps increases educational attainment, reduces criminal activity, and increases employment and earnings (at least for some years after the program).

Flores and Flores-Lagunes (2009) appear to be the first to assess the causal mechanisms of the program and find a positive direct effect on earnings after controlling for the mediator work experience which they assume to be conditionally exogenous given pre-treatment variables. Flores and Flores-Lagunes (2010) bound the indirect effects of Job Corps on employment and earnings which are mediated by the achievement of a GED, high school degree or vocational degree, as well as the direct effects based on a partial identification approach that allows for mediator endogeneity. In contrast to these studies which are concerned with labor market outcomes, we focus on the program's effects on general health. To be precise, we consider a binary health indicator (Y) evaluated 2.5 years after randomization, which is equal to one if self-assessed general health is stated to be very good and zero otherwise. In this context, employment appears to be an interesting mediator, as it is affected by Job Corps and may itself have an impact on health.

In line with this idea, Huber *et al*. (2011) find that entering employment increases self-assessed mental health when investigating a sample of German welfare recipients. Furthermore, several studies in medicine and social sciences conclude that there is a negative association between unemployment and health; see, for instance, the surveys by Jin *et al*. (1997), Björklund and Eriksson (1998) and Mathers and Schofield (1998). We therefore disentangle the total health effect into a direct and an indirect component that is due to a change in the likelihood to work. If there existed a positive total effect which, however, only operated through employment, this could imply that health care and health education were less decisive for general health than the human-capital related interventions of Job Corps which affect employability. In this context, the analysis of causal mechanisms may help to assess the usefulness of different components of a program in place.

We define employment in the first half of the second year after randomization (i.e. half way between the treatment assignment and the measurement of the outcome) as our mediator of interest (M). That is, $M = 1$ in the case of any kind of employment and $M = 0$ otherwise. We argue that the covariates to be controlled for should include potential confounders that are measured shortly before the mediator, as they may change over time, in particular as a function of the treatment. In contrast to Flores and Flores-Lagunes (2009), we therefore do not exclusively rely on pre-treatment covariates but also use variables measured in the year after treatment assignment, just before the assessment of the mediator. Nevertheless, we also condition on a rich set of pre-treatment variables, not only to control for mediator endogeneity but also to avoid confounding of the treatment induced by conditioning on post-treatment variables only.

The empirical literature (see, for instance, Mulatu and Schooler, 2002, and Llena-Nozal *et al*., 2004, among many others) suggests that socio-economic factors such as education, age and income are strongly correlated with health while they also determine an individual's employment perspectives. As discussed in Huber *et al*. (2011), similar arguments are likely to hold for the labor market history.

[13] Note that compliance with the treatment assignment was not perfect. According to Schochet *et al*. (2008) only 73% of eligible individuals actually enrolled at Job Corps centers. Here, we abstract from this issue and consider the assignment as treatment variable. Strictly speaking, we therefore consider (direct and indirect) 'intention to treat' effects rather than treatment effects.

[14] In particular, Schochet *et al*. (2001) report that the randomization of the program was successful: of 94 observed pre-treatment covariates, only five where statistically significantly different across treatment groups at the 5% level, which is what one would expect by chance.

DOI: 10.1002/jae

For example, previous jobs might have a positive or negative effect on health depending on an individual's level of stress, willingness/reluctance to work or physical strain. Furthermore, as acknowledged in Böckerman and Ilmakunnas (2009), it appears important to condition on initial (in our case: pre-mediator) health, which allows controlling for time-constant unobservable confounders. In the data, we do not only observe initial health but also health behavior prior to the mediator period such as alcohol and drug abuse.

We analyze the direct and indirect effects of the program separately by gender, in order to account for potential effect heterogeneity. We restrict the initial dataset (14,327 youths with completed baseline survey prior to the treatment assignment) to the 4352 females and 5673 males for which the post-treatment variables M and Y are observed in the follow-up survey after 2.5 years. Table V presents descriptive evidence that the selection into the mediator is indeed non-random for females and males in our evaluation sample. Individuals entering employment 1–1.5 years after randomization are on average slightly older (in the case of females), more educated (in the case of males), less often arrested, more likely to be white, less likely to receive public housing, transfer payments and food stamps, and living in smaller households at assignment. The association with household income is non-monotonic, whereas the number of children is (as expected) negatively associated with female employment and positively with male employment. Concerning the labor market history, we see a strong positive correlation between previous employment and the mediator and a negative association of the latter with being in a training activity in the year before the mediator assessment, pointing to locking-in effects. In contrast, pre-mediator health is not strongly correlated with the mediator employment. Both the differences in general health (evaluated on a scale) and the incidence of physical or emotional problems (dummy variable) are insignificant. Maybe surprisingly, alcohol abuse is higher among the working than among the non-working, while differences in illegal drug use are mostly insignificant.

In the estimation of the direct effects as well as the total and partial indirect effects based on Assumptions 3–5, we control for both the pre-assignment and post-assignment (but pre-mediator) values of these potential confounders in separate propensity score specifications (probit) for females and males.[15] We test the models of $\Pr(D=1|M,W,X)$ by the nonparametric specification test for propensity score models proposed by Shaikh *et al.* (2009), which does not reject the specifications at any conventional significance level.[16] Furthermore, we also estimate the indirect effects based on Assumptions 1 and 2, where we only condition on the pre-assignment covariates when estimating the propensity score $\Pr(D=1|M,X)$.

Table VI presents the estimated effects on females and males. The second column gives the ATE, i.e. the mean difference between treated and non-treated outcomes, along with the standard error (SE) and p-value. Taking a look at the females, the estimate suggests that Job Corps increases the incidence of a very good general health state by 2.8% points.[17] The direct effects under treatment (column 3) and non-treatment (column 4) are almost identical to the ATE and significant at the 10% level. Therefore, the program appears to have a sizable effect that is not mediated by employment. In contrast, the total and partial indirect effects based on Assumptions 1–3 (columns 5–8) as well as the indirect effects

[15] The distributions of the propensity scores across treatment states and gender are provided in Appendix A.1 (supporting information).

[16] Shaikh *et al.* (2009) show that $f_{\Pr(D=1|M,W,X)|D=1}(\rho|D=1) = \frac{\Pr(D=0)}{\Pr(D=1)}\frac{\rho}{1-\rho} f_{\Pr(D=1|M,W,X)|D=0}(\rho|D=0) \textit{for all} \rho \in (0,1)$, with $f_{Pr(D=1|M,W,X)|D=d}(\cdot|D=d)$ being the probability density function of $\Pr(D=1|M,W,X)$ conditional on $D=d$, is a testable implication of a correctly specified propensity score. Using their test based on kernel density estimation, where the bandwidth is chosen according to the Silverman (1986) rule of thumb, we obtain p-values of 0.657 and 0.431 for the propensity scores in the female and male samples, respectively. The non-rejection of the models is insensitive to using twice or half the bandwidth.

[17] The mean outcome is 0.343 among the treated and 0.315 among the non-treated such that the ATE amounts to roughly 8–9% of the mean outcomes.

DOI: 10.1002/jae

Table V. Descriptives

Variable	Females				Males			
	$M=1$	$M=0$	Diff.	p-val.	$M=1$	$M=0$	Diff.	p-val.
Socio-economic factors								
Age at assignment	18.750	18.377	0.373	0.000	18.506	17.880	0.627	0.000
Years of education at ass.	10.523	10.047	0.476	0.000	10.242	10.116	0.126	0.397
In school in yr. before ass.	0.629	0.635	−0.006	0.662	0.644	0.696	−0.052	0.000
No. of own children in household after 1st yr	0.496	0.671	−0.175	0.000	0.129	0.092	0.037	0.001
Ethnicity: black*	0.509	0.587	−0.078	0.000	0.415	0.529	−0.114	0.000
Ethnicity: white*	0.247	0.152	0.094	0.000	0.354	0.223	0.131	0.000
Household size at ass.	4.503	4.825	−0.323	0.000	4.351	4.492	−0.141	0.011
Low household income in yr before ass.**	0.313	0.382	−0.069	0.000	0.227	0.288	−0.062	0.000
High household income in yr before ass.**	0.329	0.393	−0.063	0.000	0.366	0.405	−0.040	0.003
No. of times in welfare before ass.	2.079	2.326	−0.247	0.000	1.915	2.134	−0.218	0.000
Food stamps in yr before ass.	0.506	0.593	−0.086	0.000	0.337	0.423	−0.086	0.000
Public assistance in yr before ass.	0.251	0.283	−0.032	0.020	0.240	0.262	−0.022	0.067
In public housing 1 yr after ass.	0.164	0.231	−0.067	0.000	0.118	0.171	−0.053	0.000
Transfer payments in 1st yr after ass.	0.507	0.621	−0.114	0.000	0.276	0.376	−0.099	0.000
Ever arrested before ass.	0.160	0.172	−0.013	0.276	0.317	0.329	−0.012	0.357
No. of arrests in 1st yr after ass.	0.071	0.056	0.015	0.123	0.242	0.340	−0.098	0.000
Pre-mediator labor market state								
Ever worked before ass.	0.132	0.183	−0.051	0.000	0.123	0.155	−0.032	0.001
Worked in yr before ass.	0.707	0.480	0.227	0.000	0.728	0.526	0.202	0.000
Worked in 1st yr after ass.	0.812	0.389	0.423	0.000	0.828	0.424	0.404	0.000
Worked some time in months 9–12 after ass.	0.720	0.200	0.520	0.000	0.743	0.225	0.518	0.000
Worked all the time in months 9–12	0.384	0.009	0.375	0.000	0.394	0.014	0.381	0.000
In training in yr. before ass.	0.017	0.017	−0.000	0.903	0.019	0.019	0.000	0.911
Vocational training in months 9–12	0.208	0.257	−0.048	0.000	0.186	0.231	−0.045	0.000
Academic training in months 9–12	0.323	0.411	−0.089	0.000	0.317	0.430	−0.112	0.000
Pre-mediator health (behavior)								
Health at ass. (1 = very good, 4 = bad)	1.721	1.762	−0.041	0.074	1.619	1.648	−0.029	0.143
Health after 1 year (1 = very good, 4 = bad)	1.847	1.868	−0.020	0.387	1.721	1.747	−0.026	0.213
Phys./emot. problems at ass.	0.055	0.056	−0.001	0.852	0.045	0.049	−0.004	0.468
Phys./emot. problems 1 yr after ass.	0.157	0.140	0.017	0.115	0.126	0.114	0.012	0.200
Alcohol abuse before ass.	0.555	0.460	0.094	0.000	0.654	0.553	0.101	0.000
Alcohol abuse 1 yr after ass.	0.239	0.159	0.080	0.000	0.366	0.245	0.121	0.000
Illegal drugs before ass.	0.004	0.004	0.000	0.998	0.005	0.006	−0.001	0.664
Illegal drugs 1 yr after ass.	0.010	0.008	0.002	0.413	0.020	0.013	0.007	0.053
Very good health after 2.5 yrs (Y)	0.333	0.333	0.000	0.999	0.428	0.413	0.015	0.287

Note:
*Baseline category is neither black nor white;
**baseline category is intermediate household income.

based on Assumptions 1 and 2 (columns 9 and 10) are close to zero and insignificant. Therefore, employment does not seem to mediate the effectiveness of the program in any important way. For males, the ATE amounts to 2.2% points and is significant at the 10% level.[18] In contrast to the females, however, we do not find any sizable direct effects, which points to effect heterogeneity w.r.t. gender. An interesting picture arises when looking at the indirect effects. While the total and partial indirect effects based on Assumptions 3–5 are all close to zero, estimation based on Assumptions 1 and 2 leads to partly conflicting results: $\hat{\delta}(1)$ is significantly positive,[19] which is at odds with $\hat{\delta}^t(1)$, as both

[18] The mean outcomes are 0.432 under treatment and 0.410 under non-treatment.

[19] In contrast, the direct effects $\hat{\theta}(1)$, $\hat{\theta}(0)$ remain virtually unchanged when conditioning on pre-treatment covariates only, both for males and females.

DOI: 10.1002/jae

Table VI. Effects on the incidence of very good general health after 2.5 years

	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}^t(1)$	$\hat{\delta}^t(0)$	$\hat{\delta}^p(1)$	$\hat{\delta}^p(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$
Females									
Effect	0.028	0.029	0.028	−0.000	0.002	−0.000	−0.000	0.003	−0.000
SE	0.014	0.016	0.016	0.001	0.002	0.000	0.001	0.004	0.001
p-value	0.044	0.068	0.071	0.953	0.261	0.827	0.830	0.352	0.770
Males									
Effect	0.022	0.002	0.002	0.000	0.003	0.000	0.001	0.011	0.001
SE	0.013	0.016	0.015	0.001	0.002	0.000	0.001	0.004	0.001
p-value	0.100	0.876	0.909	0.904	0.180	0.717	0.434	0.002	0.376

Note: Standard errors (SE) are estimated based on 1999 bootstrap draws.

estimators target the same parameter (the total of indirect effects). This again demonstrates the importance of carefully considering the choice of identifying assumptions.

To check the sensitivity of our results to potential attrition bias due to restricting our sample to individuals with observed post-treatment variables, we consider the response behavior in the follow-up period to be a function of the observed variables D, W, X. This corresponds to the missing at random assumption of Rubin (1976).[20] The latter allows correcting for attrition bias by weighting observations in the estimation by $R/\Pr(R=1|D,W,X)$, with R being the binary response indicator (see, for instance, Wooldridge, 2002, 2007). We estimate the response propensity $\Pr(R=1|D,W,X)$ using a probit model and find that controlling for attrition substantially decreases the precision of the estimates, but does not overthrow our results. We therefore conclude that for our sample of disadvantaged youths in the USA the health effects mediated by employment appear to be negligible. In contrast, our estimates point to a considerable direct effect of the program on the subjective health state of females.

6. CONCLUSION

This paper has demonstrated how to identify average direct and indirect effects of a binary treatment under selection on observables, (mainly) based on inverse probability weighting (IPW) using the treatment propensity score. Identification relies on the assumption of conditional exogeneity of the treatment as well as the mediator (the intermediate outcome of interest through which the indirect effect operates) given observed variables. We have considered two sets of assumptions: mediator exogeneity (i) given the treatment and covariates which are not influenced by the treatment (with the leading case being pre-treatment variables) and (ii) given the treatment and covariates which are themselves a function of the treatment. It has been shown that direct effects can be straightforwardly identified in either case, whereas the identification of indirect effects becomes more cumbersome in the latter case, which, however, appears more realistic in empirical applications. The identification issues for either set of assumptions have been demonstrated in a simulation study. Finally, we have provided an application to the experimental evaluation study of the Job Corps program. As the results are partly sensitive to the choice of assumptions, the importance of carefully considering the plausibility of the imposed identifying restrictions in the analysis of causal mechanisms cannot be overemphasized.

ACKNOWLEDGEMENTS

An earlier version of this paper was circulated under the title ‘Identifying causal mechanisms in experiments (primarily) based on inverse probability weighting’. I have benefited from comments by

[20] For a discussion of alternative forms of missingness in experiments, see Huber (2012).

DOI: 10.1002/jae

Guido Imbens, Michael Lechner, Giovanni Mellace, Andreas Steinmayr, Teppei Yamamoto, participants at 'Frontiers in the Analysis of Causal Mechanisms' (Harvard, March 2012), 'Symposium on Causality' (Jena, July 2012), the EALE Annual Meetings (Bonn, September 2012), the IAB conference 'Field Experiments in Policy Evaluation' (Nuremberg, October 2012) and the IFAU/IZA Conference on Labor Market Policy Evaluation (Bonn, October 2012), and three anonymous referees. Financial support from the Swiss National Science Foundation (grant PBSGP1_138770) is gratefully acknowledged.

REFERENCES

Albert JM. 2008. Mediation analysis via potential outcomes models. *Statistics in Medicine* **27**: 1282–1304.

Albert JM, Nelson S. 2011. Generalized causal mediation analysis. *Biometrics* **67**: 1028–1038.

Avin C, Shpitser I, Pearl J. 2005. Identifiability of path-specific effects. In *IJCAI-05, Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh; 357–363.

Baron RM, Kenny DA. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51**: 1173–1182.

Björklund A, Eriksson T. 1998. Unemployment and mental health: evidence from research in the Nordic countries. *Scandinavian Journal of Social Welfare* **7**: 219–235.

Böckerman P, Ilmakunnas P. 2009. Unemployment and self-assessed health: evidence from panel data. *Health Economics* **18**: 161–179.

Busso M, DiNardo J, McCrary J. 2009a. Finite sample properties of semiparametric estimators of average treatment effects. Working paper, University of Michigan.

Busso M, DiNardo J, McCrary J. 2009b. New evidence on the finite sample properties of propensity score matching and reweighting estimators. IZA Discussion Paper No. 3998.

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

Crump RK, Hotz VJ, Imbens GW, Mitnik OA. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**: 187–199.

Flores CA, Flores-Lagunes A. 2009. Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. IZA Discussion Paper No. 4237.

Flores CA, Flores-Lagunes A. 2010. Nonparametric partial identification of causal net and mechanism average treatment effects. Working paper, University of Florida.

Frölich M. 2004. Finite sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics* **86**: 77–90.

Heckman J, Navarro-Lozano S. 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics* **86**: 30–57.

Hirano K, Imbens GW, Ridder G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**: 1161–1189.

Horvitz D, Thompson D. 1952. A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association* **47**: 663–685.

Huber M. 2012. Identification of average treatment effects in social experiments under alternative forms of attrition. *Journal of Educational and Behavioral Statistics* **37**: 443–474.

Huber M, Lechner M, Wunsch C. 2011. Does leaving welfare improve health? Evidence for Germany. *Health Economics* **20** 484–504.

Huber M, Lechner M, Wunsch C. 2013. The performance of estimators based on the propensity score. *Journal of Econometrics* **175**: 1–21.

Imai K, Yamamoto T. 2013. Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Political Analysis* **21**: 141–171.

Imai K, Keele L, Yamamoto T. 2010. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* **25**: 51–71.

Imai K, Keele L, Tingley D, Yamamoto T. 2011. Unpacking the black box: learning about causal mechanisms from experimental and observational studies. *Political Science Review* **105**: 765–789.

Imai K, Tingley D, Yamamoto T. 2012. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society, Series A* **176**: 5–51.

DOI: 10.1002/jae

Imbens GW. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* **86**: 4–29.
Jin RL, Shah CP, Svoboda TJ. 1997. The impact of unemployment on health: a review of the evidence. *Journal of Public Health Policy* **18**: 275–301.
Judd CM, Kenny DA. 1981. Process analysis: estimating mediation in treatment evaluations. *Evaluation Review* **5**: 602–619.
Kang JDY, Schafer JL. 2007. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**: 523–539.
Kaufman S, Kaufman J, MacLenose R, Greenland S, Poole C. 2005. Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine* **24**: 1683–1702.
Khan S, Tamer E. 2010. Irregular identification, support conditions, and inverse weight estimation. *Econometrica* **78**: 2021–2042.
Llena-Nozal A, Lindeboom M, Portrait F. 2004. The effect of work on mental health: does occupation matter? *Health Economics* **13**: 1045–1062.
MacKinnon DP. 2008. Introduction to Statistical Mediation Analysis. Taylor & Francis: New York.
Mathers CD, Schofield DJ. 1998. The health consequences of unemployment: the evidence. *Medical Journal of Australia* **168**: 178–182.
Mulatu S, Schooler C. 2002. Causal connections between socio-economic status and health: reciprocal effects and mediating mechanisms. *Journal of Health and Social Behavior* **43**: 22–41.
Newey WK. 1984. A method of moments interpretation of sequential estimators. *Economics Letters* **14**: 201–206.
Pearl J. 2001. Direct and indirect effects. In Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufman: San Francisco; 411–420.
Pearl J. 2011. The causal mediation formula: a practitioner guide to the assessment of causal pathways. Technical Report R-379, University of California, Los Angeles.
Petersen ML, Sinisi SE, van der Laan MJ. 2006. Estimation of direct causal effects. *Epidemiology* **17**: 276–284.
Robins JM. 1986. A new approach to causal inference in mortality studies with sustained exposure periods: application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**: 1393–1512.
Robins JM. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. In Highly Structured Stochastic Systems, Green P, Hjort N, Richardson S (eds). Oxford University Press: Oxford; 70–81.
Robins JM, Greenland S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**: 143–155.
Robins JM, Richardson T. 2010. Alternative graphical causal models and the identification of direct effects. Working Paper no. 100, Center for Statistics and the Social Sciences, University of Washington.
Rosenbaum P. 1984. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of Royal Statistical Society, Series A* **147**: 656–666.
Rosenbaum P, Rubin D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55.
Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**: 688–701.
Rubin DB. 1976. Inference and missing data. *Biometrika* **63**: 581–592.
Rubin DB. 2004. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31**: 161–170.
Schochet PZ, Burghardt J, Glazerman S. 2001. National Job Corps study: the impacts of job corps on participants employment and related outcomes. Report, Mathematica Policy Research, Washington, DC.
Schochet PZ, Burghardt J, McConnell S. 2008. Does Job Corps work? Impact findings from the National Job Corps study. *American Economic Review* **98**: 1864–1886.
Shaikh AM, Simonsen M, Vytlacil EJ, Yildiz N. 2009. A specification test for the propensity score using its distribution conditional on participation. *Journal of Econometrics* **151**: 33–46.
Silverman B. 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall: London.
Simonsen M, Skipper L. 2006. The costs of motherhood: an analysis using matching estimators. *Journal of Applied Econometrics* **21**: 919–934.
Sjölander A. 2009. Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine* **28**: 558–571.
Tchetgen Tchetgen EJ, Shpitser I. 2011a. Semiparametric estimation of models for natural direct and indirect effects. Harvard University Biostatistics Working Paper 129.
Tchetgen Tchetgen EJ, Shpitser I. 2011b. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. Technical report, Harvard University School of Public Health.

DOI: 10.1002/jae

Ten Have TR, Joffe MM, Lynch KG, Brown GK, Maisto SA, Beck AT. 2007. Causal mediation analyses with rank preserving models. *Biometrics* **63**: 926–934.

VanderWeele TJ. 2009. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20**: 18–26.

Waernbaum I. 2012. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine* **31**: 1572–1581.

Wooldridge J. 2002. Inverse probability weighted M-estimators for sample selection, attrition and stratification. *Portuguese Economic Journal* **1**: 141–162.

Wooldridge J. 2007. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* **141** 1281–1301.

Zheng W, van der Laan MJ. 2012. Targeted maximum likelihood estimation of natural direct effects. *International Journal of Biostatistics* **8**: 1–40.

DOI: 10.1002/jae