Nonparametric estimation of natural direct and indirect effects based on inverse probability weighting

Abstract: Using a sequential conditional independence assumption, this paper discusses fully nonparametric estimation of natural direct and indirect causal effects in causal mediation analysis based on inverse probability weighting. We propose estimators of the average indirect effect of a binary treatment, which operates through intermediate variables (or mediators) on the causal path between the treatment and the outcome, as well as the unmediated direct effect. In a first step, treatment propensity scores given the mediator and observed covariates or given covariates alone are estimated by nonparametric series logit estimation. In a second step, they are used to reweigh observations in order to estimate the effects of interest. We establish root-n consistency and asymptotic normality of this approach as well as a weighted version thereof. The latter allows evaluating effects on specific subgroups like the treated, for which we derive the asymptotic properties under estimated propensity scores. We also provide a simulation study and an application to an information intervention about male circumcisions.

Keywords: Causal mechanisms, direct effects, indirect effects, causal channels, mediation analysis, causal pathways, series logit estimation, nonparametric estimation, inverse probability weighting, propensity score.

JEL classification: C21.

1 Introduction

While a large literature in social sciences focusses on assessing the average treatment effect (ATE) of some intervention, quite frequently, also the causal mechanisms through which the effect materializes appear interesting. Gelman and Imbens (2013), for instance, argue that in many cases not only the 'effects of causes' seem relevant, but also the 'causes of effects'. When for example assessing the earnings effect of a training program, policy makers might want to know whether the total impact comes from a change in search effort, human capital, or other mediators that are themselves affected by the training. For this reason, causal mediation analysis aims at disentangling a treatment effect into the indirect effect operating through one or several mediators as well as the direct effect, net of mediation. Even under random treatment assignment, total effects can in general not be disentangled by bluntly controlling for mediators, because this likely introduces selection bias, see Robins and Greenland (1992). However, direct and indirect effects are identified under a particular sequential conditional independence assumption that assumes the exogeneity of the treatment given observed covariates and of the mediator given observed covariates and the treatment, see for instance Imai, Keele, and Yamamoto (2010). Huber (2014) shows that under this assumption, identification is obtained by weighting observations by the inverses of particular treatment propensity scores[1] and considers semiparametric estimation based on parametric propensity score models in a simulation study and an application.

This paper is the first to consider fully nonparametric estimation of natural direct and indirect effects (in the denomination of Pearl (2001)) based on inverse probability weighting (IPW), using series logit estimation for the computation of the propensity scores. The advantage of the latter approach is that it prevents inconsistency of IPW due to an incorrectly specified parametric functional form of the propensity scores. We formally show that under the sequential conditional independence assumption and particular regularity conditions, nonparametric IPW is root-n consistent and asymptotically normal. Furthermore, our estimator attains the semiparametric efficiency bounds for mediation analysis derived by Tchetgen Tchetgen and Shpitser (2012). We therefore contribute to a growing literature concerned with assessing direct and indirect effects based on conditional independence under rather flexible model

[1] Tchetgen Tchetgen (2013) derives a related result in the context of inverse odds-ratio weighting.

assumptions,[2] see for instance Pearl (2001), Petersen, Sinisi, and van der Laan (2006), Flores and Flores-Lagunes (2009), VanderWeele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen Tchetgen and Shpitser (2012), and Vansteelandt, Bekaert, and Lange (2012), among others.[3] In addition to the evaluation of these effects in the total population, we in contrast to Huber (2014) also discuss the identification and estimation of weighted direct and indirect effects. This provides a framework for evaluating causal parameters in interesting subgroups, such as the direct and indirect effects on the treated, which are explicitly considered in this paper. Also for the estimators of the weighted effects in general and the effects on the treated with estimated propensity scores in particular, we show root-n consistency and asymptotic normality.

Furthermore, we investigate the finite sample performance of nonparametric IPW in a simulation study and compare it to other estimators considered in the literature, namely estimation based on (parametrically) simulating potential mediators and outcomes using the 'mediation' package for R by Tingley, Yamamoto, Hirose, Imai, and Keele (2014) and IPW with parametric propensity scores as in Huber (2014). Finally, we apply our estimator (as well as the other methods considered in the simulations) to the experimental evaluation of Chinkhumba, Godlonton, and Thornton (2014) who investigate an information intervention about male circumcisions and HIV risk in urban Malawi. We investigate whether the information treatment affects the event/willingness of being circumcised indirectly through a change in the assessment of the relative HIV risk for circumcised and uncircumcised males (which serves as mediator), or 'directly', i.e. through other mechanisms. The results point to a small indirect effect, while the direct effect estimates are never statistically different from zero.

The remainder of this paper is organized as follows. Section 2 defines the average natural direct and indirect effects and presents the identifying assumptions. Section 3 discusses nonparametric estimation based on IPW as well as inference and shows root-n consistency and asymptotic normality under particular regularity conditions. Sections 4 extends the identification and estimation results to weighted effects and to the effects on the treated under estimated propensity scores, respectively. Section 5 provides a simulation study. Section 6 presents an application to

[2] In contrast, the seminal papers in mediation analysis of Judd and Kenny (1981) and Baron and Kenny (1986) assume linear models for both the mediator and the outcome.

[3] See also the textbooks by MacKinnon (2008), Hong (2015), and Hayes (2017) for general discussions of mediation analysis in social sciences.

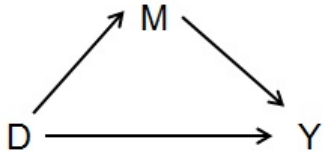
the direct and indirect effects of an information intervention about male circumcisions and HIV risk in urban Malawi. Section 7 concludes. All proofs are deferred to the Appendix.

2 Parameters of interest and identifying assumptions

2.1 Natural direct and indirect effects

We denote by D a binary intervention or treatment variable and by Y the outcome variable of interest. We would like to disentangle the causal effect of D on Y into a direct effect and an indirect impact that works through one or several discrete or continuous intermediate variables or mediators, denoted by M. Figure 1 provides a graphical illustration of the causal framework, in which arrows represent causal effects from one variable to another, but any confounders are omitted for the sake of simplicity. The definition of the total, direct, and indirect effects makes use of the potential outcome framework, see for instance Rubin (1974), and its adaptation to mediation analysis, see for instance Rubin (2004), Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007), and Albert (2008). $M(d)$, $Y(d, M(d))$ denote the potential mediator state and the potential outcome, respectively, under treatment $d \in \{0, 1\}$. $\Delta = E[Y(1, M(1)) - Y(0, M(0))]$ yields the (total) average causal effect, also known as average treatment effect (ATE), which has received much attention in the treatment or program evaluation literature, see for instance Imbens and Wooldridge (2009) for a survey.

Figure 1: Graphical illustration of the mediation framework



The (average) natural direct effect (using the denomination of Pearl (2001))[4] is defined as the mean effect of varying the treatment when keeping the mediator fixed at its potential value for

[4]Robins and Greenland (1992) and Robins (2003) refer to this parameter as the total or pure direct effect and Flores and Flores-Lagunes (2009) as net average treatment effect.

some $d \in \{0,1\}$:

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

Analogously, the (average) indirect effect is defined as the mean effect of shifting the mediator to its potential values under treatment and non-treatment when keeping the treatment fixed:

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

The ATE is the sum of the direct and indirect effects defined upon opposite treatment states:

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

The notation $\theta(1), \theta(0)$ and $\delta(1), \delta(0)$ point out that direct and indirect effects may be heterogeneous in the treatment, which allows for interaction effects between the treatment and the mediator. No effect is identifiable without assumptions, because either $Y(1,M(1))$ or $Y(0,M(0))$ (but never both) is known for any individual, while $Y(1,M(0))$ and $Y(0,M(1))$ cannot be observed for anyone (as an individual can either be treated or non-treated, but not both at the same time).

2.2 Identification

Like Imai, Keele, and Yamamoto (2010) and many others, we rely on a sequential conditional independence assumption for identification. To this end, let X denote a vector of observed covariates that potentially confound the treatment, the mediator, and the outcome. Furthermore, we denote by $\mathcal{X}, \mathcal{M}$ the supports of X and M, respectively.

Assumption 1 (conditional independence of the treatment):

$\{Y(d',m), M(d)\} \perp D | X = x$ for all $d', d \in \{0,1\}$ and $(m,x) \in \mathcal{M} \times \mathcal{X}$.

Assumption 1 requires the treatment to be conditionally independent of the potential mediator states and outcomes given X, ruling out 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. This restriction is known as conditional independence, selection on observables, or exogeneity in the treatment evaluation literature.

Assumption 2 (conditional independence of the mediator):

$Y(d', m) \perp M | D = d, X = x$ for all $d', d \in \{0, 1\}$ and $(m, x) \in \mathcal{M} \times \mathcal{X}$.

Assumption 2 requires the mediator to be conditionally independent of the potential outcomes given D and X, ruling out unobserved confounders jointly causing the mediator and the outcome conditional on the treatment and the covariates.

Assumption 3 (common support):

$\Pr(D = d | M = m, X = x)$ is bounded away from zero for all $d \in \{0, 1\}$ and $(m, x) \in \mathcal{M} \times \mathcal{X}$.

Assumption 3 is a common support restriction requiring that the conditional probability to be treated given M, X must be bounded away from zero in either treatment state. Note that $\Pr(D = d | X = x)$ must therefore be bounded away from zero on $\mathcal{X}$, too. By Bayes' theorem, Assumption 3 also implies $\Pr(M = m | D = d, X = x) > 0$ or, in the case of a continuous M, that the conditional density of M given D, X is larger than zero. Therefore, the mediator state must not be a deterministic function of the treatment conditional on X.

Identification of the natural direct and indirect effects under these or similar assumptions based on functions of the conditional mean of Y given D, M, X and the conditional density of M given D, X (the mediation formulae) has been demonstrated for instance in Pearl (2001) and Imai, Keele, and Yamamoto (2010). Huber (2014) shows that the effects as well as the mean potential outcomes may alternatively be identified by inverse probability weighting (IPW) based on the conditional probabilities $\Pr(D = 1 | M, X)$ and $\Pr(D = d | X)$, henceforth referred to as propensity scores:

$$\begin{aligned}
\theta(d) &= E\left[\left(\frac{YD}{\Pr(D=1|M,X)} - \frac{Y(1-D)}{1-\Pr(D=1|M,X)}\right)\frac{\Pr(D=d|M,X)}{\Pr(D=d|X)}\right], \\
\delta(d) &= E\left[\frac{YI\{D=d\}}{\Pr(D=d|M,X)}\left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right], \\
E[Y(d, M(d'))] &= E\left[\frac{YI\{D=d\}}{\Pr(D=d|M,X)}\frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}\right] \text{ for } d, d' \in \{1, 0\}. \qquad (4)
\end{aligned}$$

$I\{\cdot\}$ is the indicator function which is one if its argument is satisfied and zero otherwise. The expressions for $\theta(d)$ and $\delta(d)$ are (by Bayes' theorem) mathematically identical to weighting-based

representations of the direct and indirect effect relying on $\Pr(D = 1|X)$ and $\Pr(M = m|D, X)$ (rather than $\Pr(D = 1|M, X)$) suggested in Hong (2010) and Tchetgen Tchetgen and Shpitser (2012), see their 'strategy 3'. The practical advantage of the approach advocated in this paper is that $\Pr(D = 1|M, X)$ may be easier to estimate than $\Pr(M = m|D, X)$ or the respective conditional density of M in the case of a continuous M. This particularly relevant when the support of M is rich (i.e., contains many values) or M is a vector of several variables.

We note that the identification results in (4) can be generalized in various dimensions. First, replacing Y everywhere in (4) by an indicator function that Y is smaller than or equal to a particular value a, i.e. $I\{Y \leq a\}$, allows the evaluation of distributional features and effects. In this case, the expressions for $\theta(d)$ and $\delta(d)$ provide the direct and indirect effects on the cumulative distribution function (cdf) evaluated at a rather than the average effects. Likewise, the expressions for the mean potential outcomes identify the potential cdf's when Y is substituted by $I\{Y \leq a\}$. Inverting the cdf's in turn allows identifying quantile treatment effects, given that Y satisfies particular continuity conditions. See Donald and Hsu (2014) for related results in the context of (total) treatment effect evaluation. Second, the results can be extended to the evaluation of weighted direct and indirect effects as a function of the distribution of X in an analogous way as suggested in Hirano, Imbens, and Ridder (2003), henceforth HIR, in the context of the ATE. As more thoroughly discussed in Section 4, this permits the identification and (under particular assumptions) root-n-consistent estimation in specific subgroups. As one important special case of weighted IPW, Section 4.3 shows this for the treated population. Finally, the properties derived for our estimators also apply to the estimation of the direct and partial indirect effects discussed in Section 2.3. of Huber (2014) when some covariates in X are themselves affected by the treatment.

3 Estimation and inference

3.1 Nonparametric estimation

As in Huber (2014), the proposed estimators are based on normalized versions of the sample analogs of the IPW-based identification results of expression (4), with weights adding up to unity in either treatment state, see Imbens (2004) and Busso, DiNardo, and McCrary (2014). The normalized estimators of the direct effects under treatment and non-treatment, for instance,

correspond to

$$\begin{aligned}
\hat{\theta}(1) &= \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i D_i/\hat{p}(X_i)}{\frac{1}{n}\sum_{i=1}^{n} D_i/\hat{p}(X_i)} - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]}{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]}, \qquad (5)\\
\hat{\theta}(0) &= \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i D_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}{\frac{1}{n}\sum_{i=1}^{n} D_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]} - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(1-D_i)/(1-\hat{p}(X_i))}{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)/(1-\hat{p}(X_i))},
\end{aligned}$$

and the normalized estimators of the indirect effects under treatment and non-treatment correspond to

$$\begin{aligned}
\hat{\delta}(1) &= \frac{\frac{1}{n}\sum_{i=1}^{n} D_i Y_i/\hat{p}(X_i)}{\frac{1}{n}\sum_{i=1}^{n} D_i/\hat{p}(X_i)} - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i D_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}{\frac{1}{n}\sum_{i=1}^{n} D_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}, \qquad (6)\\
\hat{\delta}(0) &= \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]}{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]} - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(1-D_i)/(1-\hat{p}(X_i))}{\frac{1}{n}\sum_{i=1}^{n}(1-D_i)/(1-\hat{p}(X_i))},
\end{aligned}$$

i indexes the observations in an i.i.d. sample of size n. $\hat{p}(X_i)$ and $\hat{p}(M_i, X_i)$ denote estimates of the true propensity scores $\Pr(D = 1|X = X_i)$ and $\Pr(D = 1|M = M_i, X = X_i)$, henceforth abbreviated by $p(X_i)$ and $p(M_i, X_i)$, respectively. In contrast to Huber (2014), we estimate the propensity scores nonparametrically by series logit estimation (SLE) based on power series as in HIR. Normalized estimators for the indirect effects, denoted by $\hat{\delta}(1), \hat{\delta}(0)$, are obtained in an analogous way. See also Appendix A.1 for the normalized estimators of the potential outcomes.

To illustrate the SLE approach consider, for instance, $\hat{p}(X_i)$ and suppose that X contains only continuous variables with dimension d_x. Let $\lambda = (\lambda_1, ..., \lambda_{d_x})' \in \mathbb{Z}_+^{d_x}$ be a d_x-dimensional vector of non-negative integers where $\mathbb{Z}_+$ denotes the set of non-negative integers, and define the norm for λ as $|\lambda| = \sum_{j=1}^{d_x} \lambda_j$. Furthermore, let $\{\lambda(k)\}_{k=1}^{\infty}$ be a sequence including all distinct $\lambda \in \mathbb{Z}_+^{d_x}$ such that $|\lambda(k)|$ is non-decreasing in k and let $x^\lambda = \prod_{j=1}^{d_x} x_j^{\lambda_j}$. For any integer K_x, define $R^{K_x}(x) = (x^{\lambda(1)}, ..., x^{\lambda(K_x)})'$ as a vector of power functions. Denote by $\mathcal{L}(a) = \exp(a)/(1+\exp(a))$ the logistic cumulative distribution function (CDF). The SLE for $p(x)$ is defined as $\hat{p}(x) = \mathcal{L}\big(R^K(x)'\hat{\pi}_K\big)$ where

$$\hat{\pi}_{K_x} = \arg\max_{\pi_k} \frac{1}{n}\sum_{i=1}^{n} \Big(D_i \ln \mathcal{L}\big(R^{K_x}(X_i)'\pi_{K_x}\big) + (1-D_i)\ln\big(1-\mathcal{L}\big(R^{K_x}(X_i)'\pi_{K_x}\big)\big)\Big). \qquad (7)$$

The asymptotic properties of $\hat{p}(x)$ are discussed in Appendix A of HIR. $\hat{p}(m, x)$ is defined in an analogous way.

3.2 Asymptotic behaviour

To show root-n-consistency and asymptotic normality of our estimators, we subsequently introduce regularity conditions that are very much related to those in HIR.

Assumption 4 (distribution of (X, M)):

(i) The distribution of the $(d_m + d_x)$-dimensional vector (M, X) is absolutely continuous with probability density $f(m, x)$; (ii) $\mathcal{M}$ and $\mathcal{X}$, are Cartesian products of compact intervals; (iii) $f(m, x)$ is twice continuously differentiable, bounded above, and bounded away from 0 on $\mathcal{M} \times \mathcal{X}$.

Assumption 4 rules out that M and/or X contain any binary or discrete variables, which seems restrictive for empirical applications. We impose this assumption for the sake of ease of discussion and note that our results could be easily extended to include discrete covariates and mediators as in Donald, Hsu, and Lieli (2014). See Section 3.4 for more discussion on this matter.

Assumption 5 (smoothness of propensity scores):

(i) $p(x)$ is continuously differentiable of order $\bar{p}_x \geq 7d_x$; (ii) $p(m, x)$ is continuously differentiable of order $\bar{p}_m \geq 7(d_m + d_x)$.

Assumption 5 is analogous to Assumption 4 of HIR and requires the propensity scores to be sufficiently smooth.

Assumption 5 requires that $p(x)$ and $p(m, x)$ are continuously differentiable of order $\bar{p}_x \geq 7d_x$ and $\bar{p}_m \geq 7(d_m + d_x)$, respectively. When the numbers of continuous variables in X and M are large, this condition may look restrictive in practice, albeit holds in studies assuming parametric specifications of the propensity score functions such as probit or logit models. On the other hand, if the smoothness condition is a concern in empirical studies, then one can relax it condition by using local polynomial estimators as in Ichimura and Linton (2005) and Donald and Lieli (2014) or higher order kernel estimators as in Abrevaya, Hsu, and Lieli (2015). However, a disadvantage of these methods is that the resulting estimated propensity score functions are not necessarily bounded away from 0 and 1 in finite samples, such that trimming might be required.

Assumption 6 (series estimator):

(i) The SLE of $p(x)$ uses a power series with $K_x = n^{\nu_x}$ for some $d_x/4(\bar{p}_x - d_x) < \nu_x < 1/9$, $K_x^3 n^{-1/2} \to 0$ and $K_x^{-(\bar{p}_x + 2d_x)/2d_x} n^{1/4} \to 0$; (ii) The SLE of $p(m, x)$ uses a power series

with $K_m = n^{\nu_m}$ for some $(d_m + d_x)/4(\bar{p}_x - d_m - d_x) < \nu_m < 1/9$, $K_m^3 n^{-1/2} \to 0$ and $K_m^{-(\bar{p}_m+2d_m+2d_x)/(2d_m+2d_x)} n^{1/4} \to 0$.

Assumption 6 restricts the growth rate of the number of approximating functions to be included in the series estimator of the propensity score. Note that our assumption is stronger than Assumption 5 of HIR by requiring that $K_x^3 n^{-1/2} \to 0$ and $K^{-(\bar{p}_x+2d_x)/2d_x} n^{1/4} \to 0$, which ensures that $\sup_{x\in\mathcal{X}} |\hat{p}(x) - p(x)| = o_p(n^{-1/4})$. An analogous result applies to $p(m, x)$. These extra conditions on power series terms are needed because in contrast to HIR, our estimation approach is based on two propensity scores. When employing mean-value expansions, we require that the second order terms are of order $o_p(1)$, for which $\sup_{x\in\mathcal{X}} |\hat{p}(x) - p(x)| = o_p(n^{-1/4})$ and $\sup_{m\in\mathcal{M},x\in\mathcal{X}} |\hat{p}(m, x) - p(m, x)| = o_p(n^{-1/4})$ are sufficient conditions.

We now define several conditional moments related to potential and observed outcomes: $\rho_{dd'}(x) = E[Y(d, M(d'))|X = x]$ and $\zeta_d(m, x) = E[Y|M = m, X = x, D = d]$ for $d, d' \in \{0, 1\}$. Note that

$$\begin{aligned}
\rho_{10}(x) &= E\Big[\frac{(1-D)Y}{1-p(X)} \frac{p(M,X)}{(1-p(M,X))}\Big|X = x\Big], \\
\rho_{01}(x) &= E\Big[\frac{DY}{p(X)} \frac{1-p(M,X)}{p(M,X)}\Big|X = x\Big], \\
\zeta_1(m, x) &= E\Big[\frac{DY}{p(X,M)}\Big|M = m, X = x\Big], \\
\zeta_0(m, x) &= E\Big[\frac{(1-D)Y}{1-p(X,M)}\Big|M = m, X = x\Big].
\end{aligned}$$

Assumption 7 imposes some regularity conditions on these moments and the second moments of the potential outcomes.

Assumption 7 (moments of Y):

(i) $E[Y] < \infty$; (ii) for $d, d' \in \{0, 1\}$, $\rho_{dd'}(x)$ are continuously differentiable over $\mathcal{X}$; (iii) for $d = 0, 1$, $\zeta_d(m, x)$ is continuously differentiable over $\mathcal{M} \times \mathcal{X}$.

Under our assumptions, nonparametric IPW estimation of the direct and indirect effects using SLE-based propensity scores is root-n-consistent and asymptotically normal.

Theorem 1 *Under Assumptions 1 to 7,*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}(1)-\theta(1) \\ \hat{\theta}(0)-\theta(0) \\ \hat{\delta}(1)-\delta(1) \\ \hat{\delta}(0)-\delta(0) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0,\mathcal{V}), \tag{8}$$

where V is a 4×4 covariance matrix generated by $\psi=(\psi_{\theta(1)},\psi_{\theta(0)},\psi_{\delta(1)},\psi_{\delta(0)})'$ with

$$\begin{aligned}
\psi_{\theta(1)}(Y,M,D,X) &= \psi_{11}(Y,M,D,X)-\psi_{01}(Y,M,D,X),\\
\psi_{\theta(0)}(Y,M,D,X) &= \psi_{10}(Y,M,D,X)-\psi_{00}(Y,M,D,X),\\
\psi_{\delta(1)}(Y,M,D,X) &= \psi_{11}(Y,M,D,X)-\psi_{10}(Y,M,D,X),\\
\psi_{\delta(0)}(Y,M,D,X) &= \psi_{01}(Y,M,D,X)-\psi_{00}(Y,M,D,X),\\
\psi_{11}(Y,M,D,X) &= \frac{DY}{p(X)}-\frac{\rho_{11}(X)}{p(X)}(D-p(X))-\mu_{11},\\
\psi_{00}(Y,M,D,X) &= \frac{(1-D)Y}{1-p(X)}+\frac{\rho_{00}(X)}{1-p(X)}(D-p(X))-\mu_{00},\\
\psi_{10}(Y,M,D,X) &= \frac{DY}{p(M,X)}\frac{1-p(M,X)}{1-p(X)}+\frac{\rho_{10}(X)}{(1-p(X))}(D-p(X))\\
&\quad -\frac{\zeta_1(M,X)}{p(M,X)(1-p(X))}(D-p(M,X))-\mu_{10},\\
\psi_{01}(Y,M,D,X) &= \frac{(1-D)Y}{1-p(M,X)}\frac{p(M,X)}{p(X)}-\frac{\rho_{01}(X)}{p(X)}(D-p(X))\\
&\quad +\frac{\zeta_0(M,X)}{(1-p(M,X))p(X)}(D-p(M,X))-\mu_{01},
\end{aligned}$$

and $\mu_{dd'}=E[Y(d,M(d'))]$ for $d,d'\in\{0,1\}$.

We note that the expressions in Theorem 1 imply that our estimators attain the semiparametric efficiency bounds for mediation analysis derived in Tchetgen Tchetgen and Shpitser (2012). Finally, a relevant question for the practical implementation of the estimator is how to choose orders K_x and K_m for propensity score estimation. In the simulations presented in Section 5, we consider cross-validation for picking either parameter, see, e.g., Chapter 15 of Li and Racine (2007). We also consider overfitting by one order higher than suggested by cross-validation. Even though cross-validation w.r.t. the the functional form of the propensity score does in general not provide the optimal order for the estimation of direct and indirect effects in a given sample, the

simulation results suggest that this approach can yield satisfactory results in practice.

3.3 Inference

Inference based on the asymptotic results in Theorem 1 requires a consistent estimator of the asymptotic covariance matrix, denoted by $\mathcal{V}$. We first propose uniformly consistent estimators for $\rho_{dd'}(x)$ and $\zeta_d(m,x)$ with $d,d' \in \{0,1\}$. Let $R^{K_x}(x)$ and $R^{K_m}(m,x)$ be the column vectors of power functions used for the estimation of the propensity score functions and define

$$\hat{\rho}_{11}(x) = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\hat{p}(X_i)}R^{K_x}(X_i)'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}R^{K_x}(X_i)R^{K_x}(X_i)'\Big)^{-1}R^{K_x}(x),$$

$$\hat{\rho}_{00}(x) = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}R^{K_x}(X_i)'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}R^{K_x}(X_i)R^{K_x}(X_i)'\Big)^{-1}R^{K_x}(x),$$

$$\hat{\rho}_{10}(x) = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)}R^{K_x}(X_i)'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}R^{K_x}(X_i)R^{K_x}(X_i)'\Big)^{-1}R^{K_x}(x),$$

$$\hat{\rho}_{01}(x) = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y}{1-\hat{p}(M_i,X_i)}\frac{\hat{p}(M_i,X_i)}{\hat{p}(X_i)}R^{K_x}(X_i)'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}R^{K_x}(X_i)R^{K_x}(X_i)'\Big)^{-1}R^{K_x}(x),$$

$$\hat{\zeta}_1(m,x) = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\hat{p}(M_i,X_i)}R^{K_m}(M_i,X_i)'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}R^{K_m}(M_i,X_i)R^{K_m}(M_i,X_i)'\Big)^{-1}R^{K_m}(m,x),$$

$$\hat{\zeta}_0(m,x) = \Big(\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\hat{p}(M_i,X_i)}R^{K_m}(M_i,X_i)'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}R^{K_m}(M_i,X_i)R^{K_m}(M_i,X_i)'\Big)^{-1}R^{K_m}(m,x).$$

This permits defining the estimated influence functions as

$$
\begin{aligned}
\hat{\psi}_{\theta(1)}(Y,M,D,X) &= \hat{\psi}_{11}(Y,M,D,X) - \hat{\psi}_{01}(Y,M,D,X),\\
\hat{\psi}_{\theta(0)}(Y,M,D,X) &= \hat{\psi}_{10}(Y,M,D,X) - \hat{\psi}_{10}(Y,M,D,X),\\
\hat{\psi}_{\delta(1)}(Y,M,D,X) &= \hat{\psi}_{01}(Y,M,D,X) - \hat{\psi}_{00}(Y,M,D,X),\\
\hat{\psi}_{\delta(0)}(Y,M,D,X) &= \hat{\psi}_{10}(Y,M,D,X) - \hat{\psi}_{00}(Y,M,D,X),\\
\hat{\psi}_{11}(Y,M,D,X) &= \frac{DY}{\hat{p}(X)} - \frac{\hat{\rho}_{11}(X)}{\hat{p}(X)}(D-\hat{p}(X)) - \hat{\mu}_{11},\\
\hat{\psi}_{00}(Y,M,D,X) &= \frac{(1-D)Y}{1-\hat{p}(X)} + \frac{\hat{\rho}_{00}(X)}{1-\hat{p}(X)}(D-\hat{p}(X)) - \hat{\mu}_{00},\\
\hat{\psi}_{10}(Y,M,D,X) &= \frac{DY}{\hat{p}(M,X)}\frac{1-\hat{p}(M,X)}{1-\hat{p}(X)} + \frac{\hat{\rho}_{10}(X)}{(1-\hat{p}(X))}(D-\hat{p}(X))\\
&\quad - \frac{\hat{\zeta}_1(M,X)}{\hat{p}(M,X)(1-\hat{p}(X))}(D-\hat{p}(M,X)) - \hat{\mu}_{10},\\
\hat{\psi}_{01}(Y,M,D,X) &= \frac{(1-D)Y}{1-\hat{p}(M,X)}\frac{\hat{p}(M,X)}{\hat{p}(X)} - \frac{\hat{\rho}_{01}(X)}{p(X)}(D-\hat{p}(X))\\
&\quad + \frac{\hat{\zeta}_0(M,X)}{(1-\hat{p}(M,X))\hat{p}(X)}(D-\hat{p}(M,X)) - \hat{\mu}_{01},
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\mu}_{11} &= \frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}(X_i)}, \quad \hat{\mu}_{00} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)}{1-\hat{p}(X_i)},\\
\hat{\mu}_{10} &= \frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)},\\
\hat{\mu}_{01} &= \frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\hat{p}(M_i,X_i)}\frac{\hat{p}(M_i,X_i)}{\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)}{1-\hat{p}(M_i,X_i)}\frac{\hat{p}(M_i,X_i)}{\hat{p}(X_i)}.
\end{aligned}
$$

Furthermore, define $\hat{\psi}_i = (\hat{\psi}_{\theta(1),i}, \hat{\psi}_{\theta(0),i}, \hat{\psi}_{\delta(1),i}, \hat{\psi}_{\delta(0),i})'$ with

$$
\begin{aligned}
\hat{\psi}_{\theta(1),i} &= \hat{\psi}_{11}(Y_i,M_i,D_i,X_i) - \hat{\psi}_{01}(Y_i,M_i,D_i,X_i),\\
\hat{\psi}_{\theta(0),i} &= \hat{\psi}_{10}(Y_i,M_i,D_i,X_i) - \hat{\psi}_{00}(Y_i,M_i,D_i,X_i),\\
\hat{\psi}_{\delta(1),i} &= \hat{\psi}_{11}(Y_i,M_i,D_i,X_i) - \hat{\psi}_{10}(Y_i,M_i,D_i,X_i),\\
\hat{\psi}_{\delta(0),i} &= \hat{\psi}_{01}(Y_i,M_i,D_i,X_i) - \hat{\psi}_{00}(Y_i,M_i,D_i,X_i).
\end{aligned}
$$

Finally, let $\widehat{\mathcal{V}} = n^{-1}\sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i'$. By the same arguments as in HIR, one can show that $\widehat{V} \xrightarrow{p} \mathcal{V}$. The details are therefore omitted.

We also note that given the asymptotic normality of our estimators, the bootstrap is a consistent inference method, too, and is used in our simulations (see Section (5) and application (see Section (6). To be specific, we draw bootstrap samples of size n with replacement. We then estimate the propensity scores with the same orders K_x and K_m as chosen in the original sample for effect estimation and compute the direct and indirect effects in the bootstrap samples using the bootstrap estimates of the propensity scores. Then, the limiting distribution of $\sqrt{n}(\hat{\theta}(1) - \theta(1))$, for instance, can be approximated by $\sqrt{n}(\hat{\theta}^b(1) - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^b(1))$ with superscript b indicating a bootstrap estimate and B denoting the number of bootstrap samples. This allows computing the standard error of $\hat{\theta}(1)$ by the standard deviation of $\hat{\theta}^b(1)$ across all bootstrap samples B.

3.4 Incorporating discrete covariates and mediators

Our theory only considers the case in which X and M are vectors of continuous variables. We subsequently discuss the implications of having discrete variables among the covariates and/or mediators. We to this end assume that on top of X, there exists an additional covariate $\widetilde{X}$ satisfying $\widetilde{x} \in \{0, 1\}$ to be included in the conditioning set. Even though we focus on this binary case, we note that it easily extends to more general cases with discrete variables. For estimating the propensity score $\hat{p}(x, \tilde{x})$, we stratify the sample on the values of $\widetilde{X}$ and perform SLE as outlined in equation (7) separately in the subsamples satisfying $\widetilde{X}_i = 1$ and $\widetilde{X}_i = 0$, respectively, for all $i \in 1, ..., n$. Estimation of $\hat{p}(m, x, \tilde{x})$ proceeds analogously. Finally, to obtain $\hat{\theta}(d)$ and $\hat{\delta}(d)$ for $d \in \{0, 1\}$, we replace $\hat{p}(x)$ and $\hat{p}(m, x, \tilde{x})$ by $\hat{p}(x, \tilde{x})$ and $\hat{p}(m, x, \tilde{x})$ in (5) and (6), respectively. Concerning the regularity conditions, we note that while $\tilde{X}$ needs to be added to the conditioning set, d_x and d_m continue to denote the numbers of continuous variables in X and M, respectively. Therefore, Assumption 4, for instance, remains valid, as well as all the other results required for asymptotic normality.

However, stratifying on all possible combinations of discrete variables may in the case of a large number of strata entail the practical issue that the numbers of observations in some strata are very small (or zero) even for large n. In such cases, functional form restrictions on SLE might be imposed as in Donald, Hsu, and Lieli (2014) to mitigate such issues related to the curse of dimensionality. Firstly, one approach is to allow lower order terms in $R^{K_x}(X)$ of equation (7) – i.e. up to some lower order $L < K_x$ – to vary across strata, while higher order terms are assumed

to be homogenous across strata (and thus to be estimated in the total sample). Secondly, one may rule out specific interactions between discrete variables to reduce the dimension of strata. To see this, suppose that the set of covariates is $(X', \tilde{X}_1, \tilde{X}_2)$ where $\tilde{X}_1$ and $\tilde{X}_2$ are two binary variable, implying four strata defined by the values of $\tilde{X}_1$ and $\tilde{X}_2$. One may implement the SLE separately for subsamples with $\tilde{X}_{1i} = \tilde{X}_{2i} = 0$, $\tilde{X}_{1i} = 1$, and $\tilde{X}_{2i} = 1$, while not considering the finer strata $\tilde{X}_{1i} = \tilde{X}_{2i} = 1$, $(\tilde{X}_{1i} = 0, \tilde{X}_{2i} = 1)$, $(\tilde{X}_{1i} = 1, \tilde{X}_{2i} = 0)$. This rules out interactions between $\tilde{X}_1$ and $\tilde{X}_2$ w.r.t. their effect on the probability of $D = 1$. These two types of restrictions may be combined and do not affect the asymptotic theory if the restrictions are removed as n grows large.

4 Weighted direct and indirect effects

4.1 Identification of weighted effects

In this section, we discuss the identification, estimation, and asymptotic results for weighted effects. Let to this end $g(X)$ denote a weighting function that depends on X or subsets thereof and satisfies $|g(X)| < \infty$ and $E[g(X)] > 0$. Weighted direct and indirect effects as well as mean potential outcomes (denoted as $\theta_g(d)$, $\delta_g(d)$, $E_g[Y(d, M(d'))]$) are identified by including $g(X)/E[g(X)]$ in the respective expectation operators presented in (4):

$$\begin{aligned}
\theta_g(d) &= E\left[\frac{g(X)}{E[g(X)]}\left(\frac{YD}{\Pr(D=1|M,X)} - \frac{Y(1-D)}{1-\Pr(D=1|M,X)}\right)\frac{\Pr(D=d|M,X)}{\Pr(D=d|X)}\right], \\
\delta_g(d) &= E\left[\frac{g(X)}{E[g(X)]}\frac{YI\{D=d\}}{\Pr(D=d|M,X)}\left(\frac{\Pr(D=1|M,X)}{\Pr(D=1|X)} - \frac{1-\Pr(D=1|M,X)}{1-\Pr(D=1|X)}\right)\right]. \\
E_g[Y(d, M(d'))] &= E\left[\frac{g(X)}{E[g(X)]}\frac{YI\{D=d\}}{\Pr(D=d|M,X)}\frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}\right] \text{ for } d, d' \in \{1, 0\}. \qquad (9)
\end{aligned}$$

This allows identifying the effects for specific subgroups of interest. For instance, setting $g(X) = p(X)$ with $E[g(X)] = \Pr(D = 1)$ yields the direct and indirect effects as well as the potential outcomes among the treated. To see this, consider $E_g[Y(d, M(d'))] = E[Y(d, M(d'))|D = 1]$ and

note that (when M, X are continuous),

$$
\begin{aligned}
& E\left[\frac{p(X)}{E[p(X)]}\frac{YI\{D=d\}}{\Pr(D=d|M,X)}\frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}\right] \qquad (10)\\
= \; & E\left[\frac{p(X)}{\Pr(D=1)}E\left[E\left[\frac{YI\{D=d\}}{\Pr(D=d|M,X)}\middle|M=m,X=x\right]\frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}\middle|X=x\right]\right]\\
= \; & \int\frac{p(X)}{\Pr(D=1)}\int E[Y|D=d,M=m,X=x]\frac{\Pr(D=d'|M,X)}{\Pr(D=d'|X)}dF_{M|X=x}(m)dF_X(x)\\
= \; & \int\frac{p(X)}{\Pr(D=1)}\int E[Y(d,m)|D=d,M=m,X=x]dF_{M|D=d',X=x}(m)dF_X(x)\\
= \; & \int\frac{p(X)}{\Pr(D=1)}\int E[Y(d,m)|M(d')=m,X=x]dF_{M(d')|X=x}(m)dF_X(x)\\
= \; & \int E[Y(d,M(d'))|X=x]\frac{p(X)}{\Pr(D=1)}dF_X(x)\\
= \; & \int E[Y(d,M(d'))|X=x]dF_{X|D=1}(x)\\
= \; & E[Y(d,M(d'))|D=1].
\end{aligned}
$$

The first equation follows from the law of iterated expectations, the second from basic probability theory and replacing expectations by integrals, the third and sixth from Bayes' theorem, the fourth from Assumptions 1 and 2, and the fifth and seventh from integration. Analogously, the parameters for the nontreated are obtained by setting $g(X) = 1 - p(X)$. Furthermore, this approach can be used to asses effect heterogeneity w.r.t. X, e.g. by defining $g(X)$ as an indicator function that X takes particular (ranges of) values.

4.2 Estimation and asymptotics under known weighting functions

Based on (9), we propose the following estimators for a known weighting function $g(X)$:

$$
\begin{aligned}
\hat{\theta}_g(1) &= \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)Y_iD_i/\hat{p}(X_i)}{\frac{1}{n}\sum_{i=1}^n g(X_i)D_i/\hat{p}(X_i)} - \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)Y_i(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]}{\frac{1}{n}\sum_{i=1}^n g(X_i)(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]}\\
\hat{\theta}_g(0) &= \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)Y_iD_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}{\frac{1}{n}\sum_{i=1}^n g(X_i)D_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]} - \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)Y_i(1-D_i)/(1-\hat{p}(X_i))}{\frac{1}{n}\sum_{i=1}^n(1-D_i)/(1-\hat{p}(X_i))},\\
\hat{\delta}_g(1) &= \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)D_iY_i/\hat{p}(X_i)}{\frac{1}{n}\sum_{i=1}^n g(X_i)D_i/\hat{p}(X_i)} - \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)Y_iD_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}{\frac{1}{n}\sum_{i=1}^n g(X_i)D_i(1-\hat{p}(M_i,X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}, \qquad (11)\\
\hat{\delta}_g(0) &= \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)Y_i(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]}{\frac{1}{n}\sum_{i=1}^n g(X_i)(1-D_i)\hat{p}(M_i,X_i)/[(1-\hat{p}(M_i,X_i))\hat{p}(X_i)]} - \frac{\frac{1}{n}\sum_{i=1}^n g(X_i)Y_i(1-D_i)/(1-\hat{p}(X_i))}{\frac{1}{n}\sum_{i=1}^n g(X_i)(1-D_i)/(1-\hat{p}(X_i))},
\end{aligned}
$$

Under our previously discussed assumptions, nonparametric IPW estimation of the direct and indirect effects using SLE-based propensity scores are root-n-consistent and asymptotically nor-

mal.

Theorem 2 *Suppose that $|g(x)|$ in uniformaly bounded on $\mathcal{X}$ and $E[g(X)] > 0$. Under Assumptions 1 to 7,*

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_g(1) - \theta_g(1) \\ \hat{\theta}_g(0) - \theta_g(0) \\ \hat{\delta}_g(1) - \delta_g(1) \\ \hat{\delta}_g(0) - \delta_g(0) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \mathcal{V}_g) \tag{12}$$

where V_g is a 4×4 covariance matrix generated by $\psi_g = (\psi_{\theta_g(1)}, \psi_{\theta_g(0)}, \psi_{\delta_g(1)}, \psi_{\delta_g(0)})'$ with

$$\begin{aligned}
\psi_{\theta_g(1)}(Y, M, D, X) &= \psi_{11,g}(Y, M, D, X) - \psi_{01,g}(Y, M, D, X) \\
\psi_{\theta_g(0)}(Y, M, D, X) &= \psi_{10,g}(Y, M, D, X) - \psi_{00,g}(Y, M, D, X) \\
\psi_{\delta_g(1)}(Y, M, D, X) &= \psi_{11,g}(Y, M, D, X) - \psi_{10,g}(Y, M, D, X) \\
\psi_{\delta_g(0)}(Y, M, D, X) &= \psi_{01,g}(Y, M, D, X) - \psi_{00,g}(Y, M, D, X) \\
\psi_{11,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]}\Big(\frac{DY}{p(X)} - \frac{\rho_{11}(X)}{p(X)}(D - p(X)) - \mu_{11,g}\Big), \\
\psi_{00,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]}\Big(\frac{(1-D)Y}{1-p(X)} + \frac{\rho_{00}(X)}{1-p(X)}(D - p(X)) - \mu_{00,g}\Big), \\
\psi_{10,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]}\Big(\frac{DY}{p(M,X)}\frac{1-p(M,X)}{1-p(X)} + \frac{\rho_{10}(X)}{(1-p(X))}(D - p(X)) \\
&\quad - \frac{\zeta_1(M,X)}{P(M,X)(1-P(X))}(D - p(M,X)) - \mu_{10,g}\Big), \\
\psi_{01,g}(Y, M, D, X) &= \frac{g(X)}{E[g(X)]}\Big(\frac{(1-D)Y}{1-p(M,X)}\frac{p(M,X)}{p(X)} - \frac{\rho_{01}(X)}{p(X)}(D - p(X)) \\
&\quad + \frac{\zeta_0(M,X)}{(1-p(M,X))p(X)}(D - p(M,X)) - \mu_{01,g}\Big),
\end{aligned}$$

where $\mu_{dd',g} = E[g(X)Y(d, M(d'))]/E[g(X)]$ for $d, d' \in \{0, 1\}$.

Inference for weighted effects can be performed in a similar way as outlined in Section 3.3.

4.3 Effects on the treated with estimated propensity scores

In this section, we discuss the estimation and asymptotic results for the subgroup of treated. Note that if the propensity score $p(X)$ was known, the results of Section 4.2 with $g(X) = p(X)$ would immediately apply. However, practically more relevant is the case that $p(X)$ is unknown and

needs to be estimated, which is considered in this section. We denote by $\hat{\theta}_t(d)$ and $\hat{\delta}_t(d)$ estimates of the direct and indirect effects among the treated, $\theta_t(d) = E[Y(1, M(d)) - Y(0, M(d))|D = 1]$ and $\delta_t(d) = E[Y(d, M(1)) - Y(d, M(0))|D = 1]$, respectively.

The normalized sample analogs of the effects in (9) with an unknown weighting function $g(X) = p(X)$ correspond to:

$$
\begin{aligned}
\hat{\theta}_t(1) &= \frac{\frac{1}{n}\sum_{i=1}^n Y_i D_i}{\frac{1}{n}\sum_{i=1}^n D_i} - \frac{\frac{1}{n}\sum_{i=1}^n Y_i(1-D_i)\hat{p}(M_i,X_i)/(1-\hat{p}(M_i,X_i))}{\frac{1}{n}\sum_{i=1}^n (1-D_i)\hat{p}(M_i,X_i)/(1-\hat{p}(M_i,X_i))}, \qquad (13)\\
\hat{\theta}_t(0) &= \frac{\frac{1}{n}\sum_{i=1}^n Y_i D_i(1-\hat{p}(M_i,X_i))(\hat{p}(X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}{\frac{1}{n}\sum_{i=1}^n D_i(1-\hat{p}(M_i,X_i))(\hat{p}(X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}\\
&\quad - \frac{\frac{1}{n}\sum_{i=1}^n Y_i(1-D_i)(\hat{p}(X_i))/(1-\hat{p}(X_i))}{\frac{1}{n}\sum_{i=1}^n (1-D_i)(\hat{p}(X_i))/(1-\hat{p}(X_i))},\\
\hat{\delta}_t(1) &= \frac{\frac{1}{n}\sum_{i=1}^n D_i Y_i}{\frac{1}{n}\sum_{i=1}^n D_i} - \frac{\frac{1}{n}\sum_{i=1}^n Y_i D_i(1-\hat{p}(M_i,X_i))(\hat{p}(X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]}{\frac{1}{n}\sum_{i=1}^n D_i(1-\hat{p}(M_i,X_i))(\hat{p}(X_i))/[\hat{p}(M_i,X_i)(1-\hat{p}(X_i))]},\\
\hat{\delta}_t(0) &= \frac{\frac{1}{n}\sum_{i=1}^n Y_i(1-D_i)\hat{p}(M_i,X_i)/(1-\hat{p}(M_i,X_i))}{\frac{1}{n}\sum_{i=1}^n (1-D_i)\hat{p}(M_i,X_i)/(1-\hat{p}(M_i,X_i))} - \frac{\frac{1}{n}\sum_{i=1}^n Y_i(1-D_i)(\hat{p}(X_i))/(1-\hat{p}(X_i))}{\frac{1}{n}\sum_{i=1}^n (1-D_i)(\hat{p}(X_i))/(1-\hat{p}(X_i))},
\end{aligned}
$$

Under the same assumptions as before, nonparametric IPW estimation of the direct and indirect effects among the treated based on SLE-based propensity scores is root-n-consistent and asymptotically normal.

Theorem 3 *Under Assumptions 1 to 7,*

$$
\sqrt{n}\begin{pmatrix} \hat{\theta}_t(1) - \theta_t(1) \\ \hat{\theta}_t(0) - \theta_t(0) \\ \hat{\delta}_t(1) - \delta_t(1) \\ \hat{\delta}_t(0) - \delta_t(0) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \mathcal{V}_t) \qquad (14)
$$

where V_t is a 4×4 covariance matrix generated by $\psi_t = (\psi_{\theta_t(1)}, \psi_{\theta_t(0)}, \psi_{\delta_t(1)}, \psi_{\delta_t(0)})'$ with

$$
\begin{aligned}
\psi_{\theta_t(1)}(Y,M,D,X) &= \psi_{11,t}(Y,M,D,X) - \psi_{01,t}(Y,M,D,X) \\
\psi_{\theta_t(0)}(Y,M,D,X) &= \psi_{10,t}(Y,M,D,X) - \psi_{00,t}(Y,M,D,X) \\
\psi_{\delta_t(1)}(Y,M,D,X) &= \psi_{11,t}(Y,M,D,X) - \psi_{10,t}(Y,M,D,X) \\
\psi_{\delta_t(0)}(Y,M,D,X) &= \psi_{01,t}(Y,M,D,X) - \psi_{00,t}(Y,M,D,X) \\
\psi_{11,t}(Y,M,D,X) &= \frac{1}{E[p(X)]}\Big(D(Y-\mu_{11,t})\Big), \\
\psi_{00,t}(Y,M,D,X) &= \frac{1}{E[p(X)]}\Big(\frac{p(X)(1-D)(Y-\rho_{00}(X))}{1-p(X)} + (\rho_{00}(X)-\mu_{00,t})D\Big), \\
\psi_{10,t}(Y,M,D,X) &= \frac{1}{E[p(X)]}\Big(\frac{p(X)DY}{p(M,X)}\frac{1-p(M,X)}{1-p(X)} + \frac{\rho_{10}(X)}{1-p(X)}(D-p(X)) \\
&\quad - \frac{p(X)\zeta_1(M,X)}{p(M,X)(1-p(X))}(D-p(M,X)) - \mu_{10,t}D\Big), \\
\psi_{01,t}(Y,M,D,X) &= \frac{1}{E[p(X)]}\Big(\frac{(1-D)Y}{1-p(M,X)}p(M,X) + \frac{\zeta_0(M,X)}{1-p(M,X)}(D-p(M,X)) - \mu_{01,t}D\Big),
\end{aligned}
$$

where $\mu_{dd',t} = E[p(X)Y(d,M(d'))]/E[p(X)]$ for $d,d' \in \{0,1\}$.

5 Simulations

This section presents a brief simulation study in which we investigate the finite sample performance of nonparametric IPW based on (5) and SLE of the propensity scores, as well as of alternative estimators by considering the following data generating process:

$$
\begin{aligned}
D &= I\{\beta(X_1^2 + X_2) + \epsilon_D > 0\}, \quad M = \beta(D + X_1^2 + X_2) + \epsilon_M, \\
Y &= D + M + \beta[(1+D)(X_1^2 + X_2) + DM(1 + X_1^2 + X_2)] + \epsilon_Y, \qquad (15) \\
&\quad \text{with } X_1, \epsilon_D, \epsilon_M, \epsilon_Y \sim \mathcal{N}(0,1), \quad X_2 \sim \text{binomial}(0.5), \text{ independently of each other.}
\end{aligned}
$$

X_1 and X_2 are observed covariates and follow standard normal and binomial distributions, respectively. We note that X_1 enters the equations of the continuous outcome Y and mediator M as well as the index function of the binary treatment D both linearly and quadratically. ϵ_D, ϵ_M, ϵ_Y are random and standard normally distributed unobservables. The parameter β gauges the degree of confounding, i.e. how strongly the covariates jointly affect D, M, and Y. β also determines the level of effect heterogeneity across values of the mediator and the covariates rooted

Table 1: Simulation results for $\beta = 0.1, 0.2$ and $n = 1000, 4000$

$n = 1000$	$\theta(1)$				$\theta(0)$				$\delta(1)$				$\delta(0)$			
$\beta = 0.1$	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov
ipw cv	0.01	0.07	0.07	0.95	0.06	0.10	0.11	0.93	-0.05	0.11	0.12	0.93	-0.01	0.07	0.07	0.95
ipw ofit	-0.00	0.07	0.07	0.96	0.01	0.07	0.07	0.96	-0.01	0.08	0.09	0.96	-0.00	0.07	0.07	0.96
sem ipw	0.05	0.07	0.09	0.89	0.04	0.07	0.08	0.90	0.04	0.09	0.10	0.92	0.03	0.07	0.07	0.93
sim	0.05	0.07	0.08	0.88	0.03	0.07	0.08	0.91	0.04	0.09	0.10	0.91	0.03	0.07	0.07	0.92
$\beta = 0.2$	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov
ipw cv	-0.01	0.08	0.09	0.97	0.03	0.13	0.13	0.95	-0.04	0.13	0.14	0.94	-0.01	0.07	0.07	0.95
ipw ofit	-0.03	0.08	0.09	0.97	-0.01	0.09	0.10	0.98	-0.00	0.12	0.12	0.96	-0.01	0.08	0.08	0.96
sem ipw	0.21	0.09	0.22	0.38	0.06	0.09	0.11	0.89	0.25	0.13	0.28	0.49	0.10	0.07	0.13	0.71
sim	0.19	0.09	0.21	0.40	0.04	0.09	0.10	0.92	0.26	0.14	0.29	0.46	0.11	0.07	0.13	0.63
$n = 4000$	$\theta(1)$				$\theta(0)$				$\delta(1)$				$\delta(0)$			
$\beta = 0.1$	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov
ipw cv	0.00	0.03	0.03	0.96	0.00	0.03	0.03	0.96	0.00	0.04	0.04	0.95	0.00	0.03	0.03	0.95
ipw ofit	0.00	0.03	0.03	0.96	0.00	0.03	0.03	0.97	0.00	0.04	0.04	0.96	0.00	0.03	0.03	0.96
sem ipw	0.05	0.03	0.06	0.66	0.04	0.03	0.05	0.84	0.05	0.04	0.07	0.76	0.03	0.03	0.05	0.81
sim	0.04	0.03	0.05	0.65	0.02	0.03	0.04	0.83	0.05	0.04	0.07	0.75	0.03	0.03	0.05	0.79
$\beta = 0.2$	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov	bias	sd	rm	cov
ipw cv	-0.01	0.04	0.04	0.95	0.00	0.04	0.04	0.98	-0.00	0.05	0.05	0.95	0.00	0.04	0.04	0.95
ipw ofit	-0.02	0.04	0.05	0.95	-0.01	0.04	0.04	0.97	0.00	0.06	0.06	0.96	0.00	0.04	0.04	0.96
sem ipw	0.20	0.05	0.21	0.00	0.05	0.04	0.07	0.77	0.26	0.06	0.27	0.01	0.11	0.04	0.11	0.14
sim	0.16	0.04	0.16	0.00	0.00	0.04	0.04	0.88	0.27	0.07	0.28	0.01	0.11	0.04	0.12	0.09

Note: 'bias', 'sd', 'rm', and 'cov' denote the bias, standard deviation, root mean squared error, and coverage rate of the true effect when bootstrapping/simulating 199 times. 'ipw cv', 'ipw ofit', 'sem ipw', and 'sim' denote IPW using SLE based on cross-validation, IPW using SLE based on overfitting, semiparametric IPW using probit, and simulation-based estimation, respectively.

in the interaction terms between D, M, X_1, and X_2 in the outcome equation. In our simulations with 1000 replications, we set $\beta = 0.1, 0.2$, implying low and stronger confounding and effect heterogeneity, respectively, and consider two sample sizes of $n = 1000, 4000$.[5]

We investigate the performance of the following estimators: (i) Nonparametric IPW using SLE-based propensity scores as outlined in Section 3, where the orders K_x and K_m for the series approximations are chosen by leave-one-out cross-validation for either propensity score given X_1, X_2 and X_1, X_2, M, respectively. For implementation, we make use of some functions provided in the 'LARF' package of An and Wang (2016) for the statistical software R. On top of IPW based on cross-validated propensity scores (ipw cv), we also consider an overfitted version (ipw ofit) where K_x and K_m are one order higher than suggested by cross-validation. (ii) Semiparametric IPW based on parametric plug-in estimators as in Huber (2014) (sem ipw), using probit models for the treatment propensity scores. Note that all IPW estimators of (i) and (ii)

[5]For $\beta = 0.1$, the averages of D and M are 0.56 and 0.21, respectively, in our simulations. For $\beta = 0.2$, the averages of D and M are 0.61 and 0.42, respectively.

apply a trimming rule that discards observations with propensity scores smaller than 0.02 or larger than 0.98 to prevent exploding weights due to small denominators.[6] (iii) Simulation-based estimation (sim) as proposed by Tingley, Yamamoto, Hirose, Imai, and Keele (2014), see the 'mediation' package for R. The implementation considered here estimates the mediator and outcome models as linear functions of D, X_1, X_2 and D, M, DM, X_1, X_2, respectively, in order to simulate potential mediators and outcomes and compute direct and indirect effects. It therefore omits the squared terms of X_1 as well as interactions with covariates in the outcome equation.

Table 1 reports the bias, standard deviation (sd), root mean squared error (rm) of the estimators under various combinations of β and n. As a general pattern, nonparametric IPW becomes relatively more competitive when compared to IPW with probit-based propensity scores and simulation-based estimation as n and/or β increase. In any simulation design with $n = 4000$, our procedure with both cross-validated or overfitted propensity scores by and large dominates the other estimators in terms of the root mean squared error. This is driven by the lower absolute bias of nonparametric IPW due to not imposing parametric assumptions on the propensity score or mediator/outcome models (in particular when β is large), while all methods are quite comparable in terms of standard deviations. We also see that for $n = 1000$, the overfitted version of our method moderately outperforms estimation based on cross-validation. Under the larger sample size, however, both methods are very similar in terms of bias, standard deviation, and root mean squared error.

Table 1 also provides the coverage rates (cov), i.e. the share of simulations in which the 95% confidence intervals of the estimators include the respective true effect. For the IPW methods, the standard errors required for computing confidence intervals in each simulation sample rely on 199 bootstrap samples in which both the propensity scores and effects are re-estimated. Also for the method of Tingley, Yamamoto, Hirose, Imai, and Keele (2014), inference is based on 199 simulation steps. In the case of nonparametric IPW with cross-validation or overfitting, empirical coverage is generally close to the nominal level of 95% for either sample size and choice of β. In contrast, the coverage rates of the biased methods, namely IPW with probit-based propensity scores and simulation-based estimation, are in many cases substantially smaller than 95%, in

[6] Only few observations in our simulations need to be trimmed. For $(\beta = 0.1, n = 1000)$, on average 0.36 overfitted SLE-based propensity scores including both the covariates and the mediator lie outside the $[0.02, 0.98]$ interval. This is also the case for on average 0.05 cross-validated SLE-based propensity scores. For $(\beta = 0.2, n = 1000)$, the respective numbers are 2.69 and 1.51 for the overfitted and cross-validated SLE-based propensity scores, respectively. None of the probit-based propensity scores are trimmed in any simulation design.

particular for $\beta = 0.2$ and/or $n = 4000$.

6 Application

We apply our methods to experimental data from Chinkhumba, Godlonton, and Thornton (2014) who aim at measuring the demand for adult medical male circumcision among 1,634 uncircumcised men in urban Malawi as a function of randomized subsidies for circumcision and comprehensive information on circumcision and HIV. In our mediation analysis, we focus on the information campaign only, which was randomized independently of the financial subsidies. Men receiving comprehensive information were informed that circumcision is partially protective against HIV transmission (based on other empirical studies) at the baseline survey in 2010, while those who did not receive it were only told about the (circumcision) services of the experimenters' partner clinic. We are interested in the effect of information (D) on a binary outcome taking the value one if a male has already been circumcised or claimed to be willing to ever get circumcised at the follow up survey (Y) in 2011, roughly one year after the baseline survey and treatment assignment.[7] We assess the direct impact of the treatment as well as its indirect effect operating through (a change in) the risk assessment of HIV with and without circumcision. To this end, the mediator is defined as dummy for whether uncircumcised males are considered to be more prone to HIV risk than circumcised ones in the follow up survey (M). We therefore aim at evaluating whether it is HIV risk assessment or other potential causal mechanisms subsumed into the direct effect, e.g. salience about the availability of circumcision or attitudes towards circumcision, that drive any impact of the information treatment.

We control for the following baseline covariates (X) in our estimation, in particular to account for the endogeneity of the mediator, which is in contrast to the treatment not randomly assigned: age, education (measured in categories of up to 11 years of education, 12 years, or 13 years and more), a dummy for having ever had sex, a dummy for condom use at last sex as a proxy for sexual (risk) behavior, monthly expenditures in Malawian Kwacha (MKW), and dummies for having a working TV and/or stereo system as well as a working car as wealth proxies. We confine our sample to those 1,147 men without missing information in the covariates, the mediator, and the

[7] To be concise, Y is defined as one if either the interviewee stated to be circumcised or was willing to ever get circumcised, or the administrative records of the partner clinic show that he actually got circumcised, see the discussion in Chinkhumba, Godlonton, and Thornton (2014).

Table 2: Descriptives

	$D=1$		$D=0$		$M=1$		$M=0$	
	mean	std.dev	mean	std.dev	mean	std.dev	mean	std.dev
age (in years; discrete)	26.72	5.33	26.43	6.11	26.48	5.80	27.19	5.10
12 years of education (binary)	0.40	0.49	0.42	0.49	0.41	0.49	0.38	0.49
13+ years of education (binary)	0.18	0.38	0.21	0.40	0.19	0.39	0.21	0.41
ever had sex (binary)	0.87	0.34	0.88	0.32	0.88	0.33	0.85	0.35
used condom at last sex (binary)	0.38	0.48	0.42	0.49	0.41	0.49	0.31	0.46
expenditures (in 1000 MKW; continuous)	21.72	32.47	21.98	25.61	21.04	23.59	26.65	51.91
has a working tv / stereo system (binary)	0.78	0.42	0.78	0.41	0.79	0.41	0.76	0.43
has a working car (binary)	0.13	0.34	0.15	0.36	0.14	0.35	0.13	0.34
Y: got or would get circumcised (binary)	0.76	0.43	0.76	0.43	0.80	0.40	0.48	0.50

Note: Sample consists of 1,147 men without missing information in covariates, mediators, and outcomes (note that there are no missing values in the randomly assigned treatment). 'mean' and 'std.dev' denotes the mean and standard deviation, respectively. *: 'believe about HIV risk' is 0 if individual believes circumcised men to have higher HIV risk than uncircumcised men, 1 if the risk is believed to be the same for both groups, and 2 if uncircumcised men are believed to bear a higher HIV risk.

outcome (while there are no missing values in the randomly assigned treatment). Table 2 provides descriptive statistics (means and standard deviations) for the covariates and the outcome across the various treatment and mediator states. We refer to Chinkhumba, Godlonton, and Thornton (2014) for more details about the survey design, the variables, and attrition patterns.

Table 3 reports the direct and indirect effects based on the same estimators as considered in the simulations of Section 5. In the case of IPW, we again discard observations with extreme propensity scores using the 0.02 trimming rule. No observations are affected by trimming when estimating the propensity score by probit or based on SLE with cross-validation, while 2 subjects are discarded when using an overfitted propensity score where SLE uses one order higher than suggested by cross-validation. Standard errors (se) of any IPW method are based on bootstrapping the effects 1999 times[8] and p-values (pval) are computed using the t-statistic. For the method of Tingley, Yamamoto, Hirose, Imai, and Keele (2014), p-values are computed by simulating potential mediators and outcomes 1999 times, as implemented in the 'mediation' package for R.[9]

None of the direct effects is statistically different from zero at conventional levels of significance. All of the indirect effects are positive, amounting to a small increase in (intended) cir-

[8]For IPW using SLE with cross-validation for propensity score estimation, we also considered asymptotic approximations based on Theorem 1 (see Section 3.3) to compute standard errors in order to check robustness across inference methods. The obtained p-values are mostly similar to the bootstrap-based ones reported in the first line of Table 3, namely 0.90, 0.77, 0.00, and 0.07 for estimates of $\theta(1)$, $\theta(0)$, $\delta(1)$, and $\delta(0)$, respectively.

[9]Note that standard errors are not provided for the simulation-based estimators.

Table 3: Application

	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
	est	se	pval	est	se	pval	est	se	pval	est	se	pval
ipw cv	-0.00	0.02	0.89	-0.01	0.02	0.74	0.01	0.01	0.08	0.01	0.01	0.09
ipw ofit	-0.00	0.03	1.00	-0.01	0.03	0.72	0.01	0.01	0.19	0.01	0.01	0.57
semi ipw	-0.00	0.02	0.89	-0.01	0.02	0.74	0.01	0.01	0.08	0.01	0.01	0.09
sim	-0.00		0.86	-0.01		0.70	0.01		0.12	0.01		0.11

Note: 'est', 'se', and 'pval' denote the effect estimate, the standard error, and the p-value, respectively. 'ipw cv', 'ipw ofit', 'semi ipw', and 'sim' denote IPW using SLE based on cross-validation, IPW using SLE based on overfitting, semiparametric IPW using probit, and simulation-based estimation, respectively. Standard errors and/or p-values are based on 1999 bootstrap replications (in the case of IPW) or simulations (in the case of 'sim').

cumcision of roughly one percentage point. In the case of IPW using SLE with cross-validation or probit estimation of the propensity score, the indirect effects are significant at the 10% level, while p-values under simulation-based estimation (sim) are slightly higher. The p-values of IPW with overfitted propensity scores are far from any conventional level of statistical significance, but the point estimates of the indirect effects are similar to those of other procedures. All in all, the results suggest that the information intervention has a very moderate positive effect on the outcome through a change in the assessment of HIV risk with and without circumcision. In contrast, the insignificant direct effects do not point to further important causal mechanisms through which information affects (intended) circumcision.[10]

7 Conclusion

This paper proposed a fully nonparametric estimator of natural direct and indirect effects in the total population or specific subgroups based on inverse probability weighting and series logit estimation of the propensity scores when invoking a sequential conditional independence assumption. We established the conditions required for the root-n consistency and asymptotic normality of our estimator and investigated its finite sample performance in a simulation study. Finally, we applied our method to experimental data from Malawi to evaluate the direct effect of receiving information on circumcision on the inclination towards circumcision, as well as the indirect effect mediated by the risk assessment of HIV with and without circumcision. The results suggest a

[10]While the effects discussed here refer to the total population, Appendix A.4 presents IPW-based estimates of direct and indirect effects on the treated as an illustration of the weighting approach outlined in Section 4.3. Statistical power is, however, considerably lower and with one exception, none of the estimates are significant at conventional levels.

very moderate positive indirect effect while the direct effect is virtually zero and never statistically significant.

References

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015): "Estimating Conditional Average Treatment Effects," *Journal of Business & Economic Statistics*, 33(4), 485–505.

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

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

An, W., and X. Wang (2016): "Instrumental Variable Estimation of Causal Effects through Local Average Response Functions," *Journal of Statistical Software*, 71, 1–13.

Baron, R. M., and D. A. Kenny (1986): "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182.

Busso, M., J. DiNardo, and J. McCrary (2014): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *forthcoming in the Review of Economics and Statistics*.

Chinkhumba, J., S. Godlonton, and R. Thornton (2014): "The Demand for Medical Male Circumcision," *American Economic Journal: Applied Economics*, 6, 152–177.

Donald, S. G., and Y.-C. Hsu (2014): "Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models," *Journal of Econometrics*, 178, 383–397.

Donald, S. G., Y.-C. Hsu, and R. P. Lieli (2014): "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business & Economic Statistics*, 32(3), 395–415.

Donald, S. G., Y.-C. H., and R. P. Lieli (2014): "Inverse Probability Weighted Estimation of Local Average Treatment Effects: A Higher Order MSE Expansion," *Statistics and Probability Letters*, 95, 132–138.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA Dicussion Paper No. 4237*.

Gelman, A., and G. Imbens (2013): "Why ask Why? Forward Causal Inference and Reverse Causal Questions," *NBER Working Paper No. 19614*.

Hayes, A. F. (2017): *An introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press, New York, USA.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Hong, G. (2010): "Ratio of mediator probability weighting for estimating natural direct and indirect effects," in *JSM Proceedings, Biometrics Section*, pp. 2401–2415. American Statistical Association, Alexandria, VA.

Hong, G. (2015): *Causality in a social world: Moderation, mediation and spill-over*. John Wiley & Sons, Ltd., West Sussex, UK.

Hsu, Y.-C. (2016): "Consistent Tests for Conditional Treatment Effects," *working paper, Academia Sinica, Taipei*.

Huber, M. (2014): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

Ichimura, H., and O. Linton (2005): "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," in *Identification and Inference for Econometric Models: essays in honor of Thomas Rothenberg*, ed. by D. Andrews, and J. Stock. Cambridge University Press.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., and T. Yamamoto (2013): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *Political Analysis*, 21, 141–171.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

Li, Q., and J. S. Racine (2007): *Nanparametric econometrics: theory and practice*. Princeton Universiry Press, Princeton, New Jersey.

MacKinnon, D. P. (2008): *Introduction to Statistical Mediation Analysis*. Taylor and Francis, New York.

Pearl, J. (2001): "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): "Estimation of Direct Causal Effects," *Epidemiology*, 17, 276–284.

Robins, J. M. (2003): "Semantics of causal DAG models and the identification of direct and indirect effects," in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

Robins, J. M., and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143–155.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (2004): "Direct and Indirect Causal Effects via Potential Outcomes," *Scandinavian Journal of Statistics*, 31, 161–170.

Tchetgen Tchetgen, E. J. (2013): "Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis," *Statistics in Medicine*, 32, 4567–4580.

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

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

Tingley, D., T. Yamamoto, K. Hirose, K. Imai, and L. Keele (2014): "Mediation: R package for causal mediation analysis," *Journal of Statistical Software*, 59, 1–38.

VanderWeele, T. J. (2009): "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects," *Epidemiologic Methods*, 1, 129–158.

A Appendix

A.1 Proof of Theorem 1

Define $Y_{dd'} = Y(d, M(d'))$ and $\mu_{dd'} = E[Y_{dd'}]$ for $d, d' \in \{0, 1\}$. By Assumptions 1 and 2,

$$\mu_{11} = E\Big[\frac{DY}{p(X)}\Big], \quad \mu_{00} = E\Big[\frac{(1-D)Y}{1-p(X)}\Big],$$
$$\mu_{10} = E\Big[\frac{DY}{p(M,X)}\frac{1-p(M,X)}{1-p(X)}\Big], \quad \mu_{01} = E\Big[\frac{(1-D)Y}{1-p(M,X)}\frac{p(M,X)}{p(X)}\Big],$$

see equations (4) and (5) of Huber (2014). This allows defining the direct and indirect effects of interest in terms of $\mu_{dd'}$, e.g. $\theta(1) = \mu_{11} - \mu_{01}$. We estimate $\mu_{dd'}$ for $d, d' \in \{0, 1\}$ by the normalized sample analogs

$$\hat{\mu}_{11} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}(X_i)}, \quad \hat{\mu}_{00} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)}{1-\hat{p}(X_i)},$$
$$\hat{\mu}_{10} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{D_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)},$$
$$\hat{\mu}_{01} = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-\hat{p}(M_i,X_i)}\frac{\hat{p}(M_i,X_i)}{\hat{p}(X_i)}\Big/\frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)}{1-\hat{p}(M_i,X_i)}\frac{\hat{p}(M_i,X_i)}{\hat{p}(X_i)}.$$

To prove Theorem 1, it is thus sufficient to show that for $d, d' \in \{0, 1\}$,

$$\sqrt{n}(\hat{\mu}_{dd'} - \mu_{dd'}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{dd'}(Y_i, M_i, D_i, X_i) + o_p(1).$$

As the results regarding μ_{11} and μ_{00} have been established by HIR, we subsequently focus on the proof for μ_{10} and note that the derivations for μ_{01} proceed in an analogous way. Let $\tilde{\mu}_{10}$ be the numerator of $\hat{\mu}_{10}$ and $\tilde{\omega}_{10}$ be the denominator of $\hat{\mu}_{10}$. We first show that

$$\sqrt{n}(\tilde{\mu}_{10} - \mu_{10}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{10}(Y_i, M_i, D_i, X_i) + o_p(1),$$

using a similar approach as HIR. Note that

$$\begin{aligned}\sqrt{n}(\tilde{\mu}_{10}-\mu_{10}) =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{D_iY_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)}-\mu_{10}\Big\}\\ =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{D_iY_i}{p(M_i,X_i)}\frac{1-p(M_i,X_i)}{1-p(X_i)}-\mu_{10}\Big\}\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{D_iY_i}{p(M_i,X_i)}\frac{1-p(M_i,X_i)}{(1-p(X_i))^2}(\hat{p}(X_i)-p(X_i))\Big\}\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{-D_iY_i}{p^2(M_i,X_i)}\frac{1}{1-p(X_i)}(\hat{p}(M_i,X_i)-p(M_i,X_i))\Big\}+o_p(1)\end{aligned}$$

where the second equality holds by a second order mean valued expansion around $p(X_i)$ and $p(M_i,X_i)$ and the fact that $\sup_{x\in\mathcal{X}}|\hat{p}(x)-p(x)|=o_p(n^{-1/4})$ and $\sup_{m\in\mathcal{M},x\in\mathcal{X}}|\hat{p}(m,x)-p(m,x)|=o_p(n^{-1/4})$. The converge rate of $\sup_{x\in\mathcal{X}}|\hat{p}(x)-p(x)|=O_p(K_x(\sqrt{K_x/n}+K^{-\bar{p}_x/2d_x}))$ is derived in Lemma A3 of Hsu (2016) and the conditions given in this paper are sufficient for $\sup_{x\in\mathcal{X}}|\hat{p}(x)-p(x)|=o_p(n^{-1/4})$. By the same argument as in Theorem 1 of HIR, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{D_iY_i}{p(M_i,X_i)}\frac{1-p(M_i,X_i)}{(1-p(X_i))^2}(\hat{p}(X_i)-p(X_i))\Big\}\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{E\Big[\frac{D_iY_i}{p(M_i,X_i)}\frac{1-p(M_i,X_i)}{(1-p(X_i))^2}\Big|X_i\Big](D_i-p(X_i))\Big\}+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{\rho_{10}(X_i)}{(1-p(X_i))}(D_i-p(X_i))\Big\}+o_p(1).\end{aligned}$$

Similarly, it holds that

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{-D_iY_i}{p^2(M_i,X_i)}\frac{1}{1-p(X_i)}(\hat{p}(M_i,X_i)-p(M_i,X_i))\Big\}\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{E\Big[\frac{-D_iY_i}{p^2(M_i,X_i)}\frac{1}{1-p(X_i)}\Big|M_i,X_i\Big](D_i-p(M_i,X_i))\Big\}+o_p(1)\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{-\zeta_1(M_i,X_i)}{p(M_i,X_i)}\frac{1}{1-p(X_i)}(D_i-p(M_i,X_i))\Big\}+o_p(1).\end{aligned}$$

Next, one can easily show that

$$\sqrt{n}(\tilde{w}_{10}-1)=o_p(1),$$

by replacing Y_i's with 1's in the proof for $\tilde{\mu}_{10}$ and acknowledging that the influence functions for

$\tilde{w}_{10}$ are all zero. Finally, it follows that

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{10} - \mu_{10}) &= \sqrt{n}(\frac{\tilde{\mu}_{10}}{\tilde{w}_{10}} - \mu_{10}) \\ &= \sqrt{n}(\tilde{\mu}_{10} - \mu_{10}) - \mu_{10}\sqrt{n}(\tilde{w}_{10} - 1) + o_p(1) \\ &= \sqrt{n}(\tilde{\mu}_{10} - \mu_{10}) + o_p(1),\end{aligned}$$

where the second equality follows by a mean-value expansion and the last equality holds by the fact that μ_{10} is bounded and that $\sqrt{n}(\tilde{w}_{10} - 1) = o_p(1)$. This completes the proof. □

A.2 Proof of Theorem 2

Define $\mu_{dd',g} = E[g(X)Y_{dd'}]/E[g(X)]$ for $d, d' \in \{0, 1\}$. Similarly,

$$\begin{aligned}\mu_{11,g} &= \frac{E\Big[g(X)\frac{DY}{p(X)}\Big]}{E[g(X)]}, \\ \mu_{00,g} &= \frac{E\Big[g(X)\frac{(1-D)Y}{1-p(X)}\Big]}{E[g(X)]}, \\ \mu_{10,g} &= \frac{E\Big[g(X)\frac{DY}{p(M,X)}\frac{1-p(M,X)}{1-p(X)}\Big]}{E[g(X)]}, \\ \mu_{01,g} &= \frac{E\Big[g(X)\frac{(1-D)Y}{1-p(M,X)}\frac{p(M,X)}{p(X)}\Big]}{E[g(X)]}.\end{aligned}$$

We estimate $\mu_{dd',g}$ for $d, d' \in \{0, 1\}$ by the normalized sample analogs

$$\begin{aligned}\hat{\mu}_{11,g} &= \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{D_i Y_i}{\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{D_i}{\hat{p}(X_i)}, \\ \hat{\mu}_{00,g} &= \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{(1-D_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{10,g} &= \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{D_i Y_i}{\hat{p}(M_i, X_i)}\frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{D_i}{\hat{p}(M_i, X_i)}\frac{1-\hat{p}(M_i, X_i)}{1-\hat{p}(X_i)}, \\ \hat{\mu}_{01,g} &= \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{(1-D_i)Y_i}{1-\hat{p}(M_i, X_i)}\frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{(1-D_i)}{1-\hat{p}(M_i, X_i)}\frac{\hat{p}(M_i, X_i)}{\hat{p}(X_i)}.\end{aligned}$$

Define $P_g = E[g(X)]$. We would like to show that $\sqrt{n}(\hat{\mu}_{11,g}-\mu_{11,g}) = n^{-1/2}\sum_{i=1}^{n}\psi_{\mu_{11,g}}(Y_i, M_i, D_i, X_i)+ o_p(1)$. First, we can be demonstrated that

$$
\begin{aligned}
&\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{D_iY_i}{\hat{p}(X_i)} - \mu_{11,g}\cdot P_g) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{g(X_i)D_iY_i}{p(X_i)} - \frac{g(X_i)\rho_{11}(X_i)}{p(X_i)}(D_i - p(X_i)) - \mu_{11,g}\cdot P_g\Big) + o_p(1), \text{ and} \\
&\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{D_i}{\hat{p}(X_i)} - \mu_{11,g}\cdot P_g) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{g(X_i)D_i}{p(X_i)} - \frac{g(X_i)}{p(X_i)}(D_i - p(X_i)) - P_g\Big) + o_p(1) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(X_i) - P_g) + o_p(1).
\end{aligned}
$$

Then by delta method, we have

$$
\begin{aligned}
&\sqrt{n}(\hat{\mu}_{11,g} - \mu_{11,g}) \\
=&\frac{1}{P_g}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{g(X_i)D_iY_i}{p(X_i)} - \frac{g(X_i)\mu_1(X_i)}{p(X_i)}(D_i - p(X_i)) - \mu_{11,g}\cdot P_g\Big) - \frac{\mu_{11,g}}{P_g}(g(X_i) - P_g) + o_p(1) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g(X_i)}{E[g(X)]}\Big(\frac{D_iY_i}{p(X_i)} - \frac{\rho_{11}(X_i)}{p(X_i)}(D_i - p(X_i)) - \mu_{11,g}\Big) + o_p(1).
\end{aligned}
$$

This gives the result for the $\mu_{11,g}$ case. Using the same arguments, we can derive the results for $\mu_{00,g}$, $\mu_{01,g}$ and $\mu_{10,g}$, too, which is sufficient to prove Theorem 2. □

A.3 Proof of Theorem 3

Define $\mu_{dd',t} = E[p(X)Y_{dd'}]/E[p(X)]$ for $d, d' \in \{0, 1\}$. Similarly,

$$
\begin{aligned}
\mu_{11,t} &= \frac{E\Big[p(X)\frac{DY}{p(X)}\Big]}{E[p(X)]}, \\
\mu_{00,t} &= \frac{E\Big[p(X)\frac{(1-D)Y}{1-p(X)}\Big]}{E[p(X)]}, \\
\mu_{10,t} &= \frac{E\Big[p(X)\frac{DY}{p(M,X)}\frac{1-p(M,X)}{1-p(X)}\Big]}{E[p(X)]}, \\
\mu_{01,t} &= \frac{E\Big[p(X)\frac{(1-D)Y}{1-p(M,X)}\frac{p(M,X)}{p(X)}\Big]}{E[p(X)]}.
\end{aligned}
$$

We estimate $\mu_{dd',t}$ for $d, d' \in \{0,1\}$ by the normalized sample analogs

$$\begin{aligned}
\hat{\mu}_{11,t} &= \frac{1}{n}\sum_{i=1}^{n} D_i Y_i \Big/ \frac{1}{n}\sum_{i=1}^{n} D_i, \\
\hat{\mu}_{00,t} &= \frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{(1-D_i)}{1-\hat{p}(X_i)}, \\
\hat{\mu}_{10,t} &= \frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{D_i Y_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{D_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)}, \\
\hat{\mu}_{01,t} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i)Y_i}{1-\hat{p}(M_i,X_i)}\hat{p}(M_i,X_i) \Big/ \frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i)}{1-\hat{p}(M_i,X_i)}\hat{p}(M_i,X_i).
\end{aligned}$$

Define $P_t = E[p(X)]$. Note that we have

$$\begin{aligned}
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} D_i Y_i - \mu_{11,t}P_t\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i Y_i - \mu_{11,t}P_t), \\
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} - \mu_{00,t}P_t\Big) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{p(X_i)(1-D_i)Y_i}{1-p(X_i)} + \frac{\mu_0(X_i)}{1-p(X_i)}(D_i - p(X_i)) - \mu_{00,t}P_t\Big) + o_p(1), \\
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{D_i Y_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)} - \mu_{10,t}P_t\Big) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{p(X_i)D_i Y_i}{p(M_i,X_i)}\frac{1-p(M_i,X_i)}{1-p(X_i)} + \frac{\rho_{10}(X_i)}{1-p(X_i)}(D_i - p(X_i)) \\
&\qquad\qquad - \frac{p(X_i)\zeta_1(M_i,X_i)}{p(M_i,X_i)(1-p(X_i))}(D_i - p(M_i,X_i)) - \mu_{10,t}P_t\Big) + o_p(1), \text{ and} \\
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i)Y_i}{1-\hat{p}(M_i,X_i)}\hat{p}(M_i,X_i) - \mu_{01,t}P_t\Big) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\frac{(1-D_i)Y_i}{1-p(M_i,X_i)}p(M_i,X_i) + \frac{\zeta_0(M_i,X_i)}{(1-p(M_i,X_i))}(D_i - p(M_i,X_i)) - \mu_{01,t}P_t\Big) + o_p(1).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} D_i - P_t\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - P_t) + o_p(1), \\
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} - P_t\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - P_t) + o_p(1) \\
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{p}(X_i)\frac{D_i}{\hat{p}(M_i,X_i)}\frac{1-\hat{p}(M_i,X_i)}{1-\hat{p}(X_i)} - P_t\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - P_t) + o_p(1), \text{ and} \\
&\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \frac{(1-D_i)}{1-\hat{p}(M_i,X_i)}\hat{p}(M_i,X_i) - P_t\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - P_t) + o_p(1).
\end{aligned}$$

By delta method, we have

$$\begin{aligned}
&\sqrt{n}(\hat{\mu}_{11,t} - \mu_{11,t}) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{P_t} \Big(D_i Y_i - \mu_{11,t} P_t - \mu_{11,t}(D_i - P_t) \Big) + o_p(1) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{P_t} \Big(D_i (Y_i - \mu_{11,t}) \Big) + o_p(1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\sqrt{n}(\hat{\mu}_{00,t} - \mu_{00,t}) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{P_t} \Big(\frac{p(X_i)(1-D_i)Y_i}{1-p(X_i)} + \frac{\rho_{00}(X_i)}{1-p(X_i)} (D_i - p(X_i)) - \mu_{00,t} P_t - \mu_{00,t}(D_i - P_t) \Big) + o_p(1) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{P_t} \Big(\frac{p(X_i)(1-D_i)(Y_i - \rho_{00}(X_i))}{1-p(X_i)} + (\rho_{00}(X_i) - \mu_{00,t}) D_i \Big) + o_p(1).
\end{aligned}$$

Next,

$$\begin{aligned}
&\sqrt{n}(\hat{\mu}_{10,t} - \mu_{10,t}) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{P_t} \Big(\frac{p(X_i) D_i Y_i}{p(M_i, X_i)} \frac{1 - p(M_i, X_i)}{1 - p(X_i)} + \frac{\rho_{10}(X_i)}{(1-p(X_i))} (D_i - p(X_i)) \\
&\qquad - \frac{p(X_i)\zeta_1(M_i, X_i)}{p(M_i, X_i)(1-p(X_i))} (D_i - p(M_i, X_i)) - \mu_{10,t} D_i \Big) + o_p(1), \\
&\sqrt{n}(\hat{\mu}_{01,t} - \mu_{01,t}) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{1}{P_t} \Big(\frac{(1-D_i)Y_i}{1-p(M_i, X_i)} p(M_i, X_i) + \frac{\zeta_0(M_i, X_i)}{(1-p(M_i, X_i))} (D_i - p(M_i, X_i)) - \mu_{01,t} D_i \Big) + o_p(1).
\end{aligned}$$

This completes the proof. □

A.4 Effects of circumcision on the treated

Table 4: Direct and indirect effects among treated

	$\theta(1)$			$\theta(0)$			$\delta(1)$			$\delta(0)$		
	est	se	pval	est	se	pval	est	se	pval	est	se	pval
ipw cv	0.04	0.02	0.13	0.03	0.03	0.19	0.01	0.01	0.24	0.01	0.01	0.29
ipw ofit	0.03	0.03	0.34	0.01	0.03	0.74	0.02	0.01	0.04	0.00	0.01	0.78
semi ipw	0.04	0.02	0.13	0.03	0.03	0.19	0.01	0.01	0.24	0.01	0.01	0.28

Note: 'est', 'se', and 'pval' denote the effect estimate, the standard error, and the p-value, respectively. 'ipw cv', 'ipw ofit', and 'semi ipw' denote IPW using SLE based on cross-validation, IPW using SLE based on overfitting, and semiparametric IPW using probit, respectively. Standard errors and/or p-values are based on 1999 bootstrap replications.