Marginal Interventional Effects*

Xiang Zhou and Aleksei Opacic

Harvard University

June 21, 2022

Abstract

Conventional causal estimands, such as the average treatment effect (ATE), reflect how the mean outcome in a population or subpopulation would change if *all* units received treatment versus control. Real-world policy changes, however, are often incremental, changing the treatment status for only a small segment of the population who are at or near "the margin of participation." To capture this notion, two parallel lines of inquiry have developed in economics and in statistics and epidemiology that define, identify, and estimate what we call interventional effects. In this article, we bridge these two strands of literature by defining interventional effect (IE) as the per capita effect of a treatment intervention on an outcome of interest, and marginal interventional effect (MIE) as its limit when the size of the intervention approaches zero. The IE and MIE can be viewed as the unconditional counterparts of the policy-relevant treatment effect (PRTE) and marginal PRTE (MPRTE) proposed in the economics literature. However, different from PRTE and MPRTE, IE and MIE are defined without reference to a latent index model, and, as we show, can be identified either under unconfoundedness or through the use of instrumental variables. For both scenarios, we show that MIEs are typically identified without the strong positivity assumption required of the ATE, highlight several "stylized interventions" that may be of particular interest in policy analysis, discuss several parametric and semiparametric estimation strategies, and illustrate the proposed methods with an empirical example.

*Direct all correspondence to Xiang Zhou, Department of Sociology, Harvard University, 33 Kirkland Street, Cambridge MA 02138; email: xiang_zhou@fas.harvard.edu

1 Introduction

Conventional causal estimands, such as the average treatment effect (ATE) and the average treatment effect for the treated (ATT), assess how the mean outcome in a population or subpopulation would change if *all* units received treatment versus control. Real-world policy changes, however, are often incremental, changing the treatment status for only a small segment of the population who are at or near "the margin of participation" (Heckman and Vytlacil 2001; Xie 2013). In such cases, ATE or ATT may be a misleading indicator of the impact of an intervention. For example, a college outreach program may induce an additional fraction, but not all, of eligible high-school graduates to attend college. If those affected by the college outreach program benefit much more from attending college than the average high school graduate or the average college-goer in the population, then the ATE or ATT of college attendance will underestimate the effect of college attendance among those affected by the outreach program, and thus should not be used to evaluate the impact of this intervention.

To mimic real-world policy changes, a growing body of research in statistics and epidemiology has considered the so-called interventional effects (e.g., Murphy *et al.* 2001; Robins *et al.* 2004; Taubman *et al.* 2009; Díaz and van Der Laan 2012; Díaz and van der Laan 2013; Moore *et al.* 2012; Haneuse and Rotnitzky 2013; Young *et al.* 2014; Kennedy 2019; Naimi *et al.* 2020; see also Korb *et al.* 2004; Shpitser and Pearl 2006; Eberhardt and Scheines 2007; Tian 2012 from a computer science perspective). In contrast to conventional causal estimands, interventional effects characterize how the mean outcome in the population responds to a hypothetical, and often incremental, change in the treatment assignment mechanism. For example, Taubman *et al.* (2009) evaluate how the 20-year risk of coronary heart disease in a cohort of US nurses would have changed if everyone exercising less than 30 minutes a day had increased their exercise time to 30 minutes a day. This intervention is deterministic in that it specifies each unit's treatment level A^* to be $\max(30, A)$, where A denotes the current treatment level. Interventions can also be stochastic. Díaz and van Der Laan (2012), for instance, consider an intervention in which each unit's treatment level A^* is a random draw from a location-shifted version of the observed treatment distribution. More recently, for a binary treatment, Kennedy (2019) proposes an incremental propensity score intervention (IPSI) that preserves the relative odds of receiving treatment between units with different

pretreatment characteristics. In addition to its affinity to real-world policy changes, a practical advantage of Kennedy's IPSI over conventional causal estimands is that its effect can be identified without the assumption of positivity.

Interestingly, a parallel line of literature has developed in economics that aims to address similar questions. In particular, Heckman and Vytlacil (2001, 2005) propose the concept of policy-relevant treatment effect (PRTE), which is defined as the (per capita) effect of a policy change (i.e., an intervention to the treatment variable) on the mean outcome conditional on a set of pretreatment covariates X. Under a latent index model for treatment, these authors show that PRTE can be expressed as a weighted average of the marginal treatment effect (MTE), a function defined as the conditional mean of treatment effect given $X = x$ and a latent variable representing unobserved, individual-specific resistance to treatment. Moreover, they show that the weights depend only on the conditional distributions of the propensity score before and after the policy change. Carneiro *et al.* (2010, 2011) further define the marginal PRTE (MPRTE) as a directional limit of PRTE as the alternative policy under consideration approaches the baseline policy. More recently, Zhou and Xie (2019, 2020) propose a modification of Carneiro et al.'s approach, in which they recast the PRTE and MPRTE parameters as the per capita effect of a policy change and its limit, respectively, conditional on the propensity score rather than the entire vector of pretreatment covariates. In this approach, a policy change is allowed to vary in intensity between units with different baseline propensity scores.

In this article, we bridge these two strands of literature by unifying the estimands considered in Heckman and Vytlacil (2001, 2005), Carneiro *et al.* (2010, 2011), Kennedy (2019), and Zhou and Xie (2019, 2020) in the context of a binary treatment. Specifically, we define the interventional effect (IE) and marginal interventional effect (MIE) as the unconditional counterparts of PRTE and MPRTE, respectively. By not conditioning on pretreatment covariates, IE and MIE are more flexible than the original PRTE and MPRTE because they can easily accommodate interventions that vary in intensity between units with different background characteristics. And by not conditioning on the propensity score, IE and MIE are more flexible than Zhou and Xie's modified PRTE and MPRTE because they do not restrict the intensity of the intervention to be a function of the baseline propensity score. Moreover, IE can be viewed as a per capita version of the interventional effects discussed in the statistics and epidemiology literature. Both IE and MIE are defined without

reference to any causal assumptions or identification strategy; as we will show, they can be identified either under the assumption of unconfoundedness or through the use of instrumental variables (IV).

Under unconfoundedness, we show that both IE and MIE are identified as a weighted mean of the conditional average treatment effect (CATE) given pretreatment covariates, where the weight is proportional to the increment of the propensity score under the intervention or its local derivative. In particular, the effect of the IPSI considered in Kennedy (2019) can be viewed as a special case of IE. This special case is especially interesting because the corresponding MIE coincides with what Li *et al.* (2018) call the average treatment effect for the overlap population (ATO), which is also akin to the optimally weighted average treatment effect (OWATE) introduced in Crump *et al.* (2006). While the statistical properties of this estimand are well understood, its substantive interpretation has been elusive. Li *et al.* (2018), for example, characterized it vaguely as the average treatment effect among "the subpopulation that currently receives either treatment in substantial proportions" (p. 391). Our work enriches the scientific content of ATO by showing it to be a limit of Kennedy's IPSI, that is, the per capita effect of an infinitesimal intervention that preserves the relative odds of treatment between units with different pretreatment characteristics. In addition, we show that under unconfoundedness, conventional causal estimands such as ATE and ATT can also be viewed as special cases of MIE.

We then move beyond unconfoundedness and consider the identification of IE and MIE with instrumental variables. Following Heckman and Vytlacil (2001, 2005), we consider a latent index model with at least one continuous instrument. In this framework, IE can be identified as a weighted mean of MTE if the conditional support of the baseline propensity score contains the conditional support of the propensity score under intervention, akin to the case of PRTE. However, as we will see, support conditions are not required for identifying MIE. In cases where the propensity score under intervention depends only on the baseline propensity score, IE and MIE coincide with Zhou and Xie's (2019, 2020) modified PRTE and MPRTE; yet, our identification formula for MIE suggests a simpler estimator than that proposed in those papers. Finally, we note that in the absence of unobserved selection, MTE reduces to CATE, and the identification formulas for IE and MIE reduce to those derived under unconfoundedness.

The rest of the paper is organized as follows. In Section 2, we introduce the concepts of IE and

MIE, outline a minimal set of assumptions for them to be well-defined, and explicate their relationships with PRTE and MPRTE. In Section 3, we discuss the identification and estimation of MIE under the assumption of unconfoundedness, highlight several special cases and their connections to existing causal estimands, and illustrate them by reanalyzing a dataset on the clinical effects of right heart catheterization. In Section 4, we discuss the identification and estimation of MIE with instrumental variables and illustrate this approach by revisiting a study on the economic returns to college. Section 5 concludes the paper.

2 Marginal Interventional Effects

2.1 Interventional Effects

Let A denote a binary treatment and Y an outcome of interest. Throughout the paper, we assume that all random variables are defined in a sample space composed of all units in a superpopulation. Thus, for a given unit i, A_i and Y_i are viewed as fixed. In other words, the randomness of A and Y stems solely from population heterogeneity. For notational conciseness, we omit the unit subscript in most of our exposition.

Consider an intervention to the treatment assignment mechanism such that it changes the proportion of units receiving treatment in the population. Let A^* and Y^* denote the treatment and outcome variables under such an intervention. Assuming that $\mathbb{E}[A^*] \neq \mathbb{E}[A]$, we define the interventional effect (IE) as the change in the mean outcome per net person shifted into treatment:

$$\text{IE} = \frac{\mathbb{E}[Y^*] - \mathbb{E}[Y]}{\mathbb{E}[A^*] - \mathbb{E}[A]}. \tag{1}$$

In statistics and epidemiology, scholars have defined interventional effects in the form of $\mathbb{E}[Y^*] - \mathbb{E}[Y]$, without standardizing by the change in the mean treatment status (e.g., Robins *et al.* 2004; Díaz and van Der Laan 2012). Compared with the unstandardized IE, our definition is especially useful in policy contexts where the treatment is expensive such that $\mathbb{E}[A^*] - \mathbb{E}[A]$ can only take a small vaule due to budget constraints. For example, when considering the benefits of a higher education expansion, $\mathbb{E}[Y^*] - \mathbb{E}[Y]$ might achieve its highest value under a "college-for-all" policy where all eligible young adults attended college (i.e., $\mathbb{E}[A^*] = 1$), but such a dramatic expansion of

college enrollments might not be economically or politically feasible. In such cases, policy makers may want to consider interventions that yield relatively large *per-person benefit*s as defined by equation (1).

The IE defined in equation (1) is similar to Heckman and Vytlacil's (2001) (per capita) PRTE, which is defined as the change in the mean outcome per net person shifted into treatment conditional on a vector of pretreatment covariates X:

$$\text{PRTE}(x) = \frac{\mathbb{E}[Y^*|X=x] - \mathbb{E}[Y|X=x]}{\pi^*(x) - \pi(x)},$$

where $\pi(x) \triangleq \mathbb{E}[A|X=x]$ and $\pi^*(x) \triangleq \mathbb{E}[A^*|X=x]$ are the propensity scores of treatment given $X=x$ before and after the intervention.

Here, the pretreatment covariates X are assumed to be unaffected by the intervention (thus no asterisk on X). The difference between IE and PRTE(x) is important because IE is well-defined provided that the intervention changes the overall proportion of units receiving treatment. By contrast, PRTE(x) is well-defined only at the covariate values where the conditional probability of treatment is shifted. Moreover, even if the intervention changes both the overall probability of treatment and the conditional probability of treatment at all covariate values, IE $\neq \mathbb{E}[\text{PRTE}(X)]$. Instead, it is a weighted mean of PRTE(X):

$$\text{IE} = \mathbb{E}\big[\underbrace{\Big(\frac{\pi^*(X) - \pi(X)}{\mathbb{E}[\pi^*(X) - \pi(X)]}\Big)}_{\triangleq w^{\text{IE}}_{\text{PRTE}}(X)} \text{PRTE}(X)\big]. \tag{2}$$

Equation (2) makes it clear that the impact of an intervention, as measured by IE, depends not only on the interventional effects among units with the same pretreatment characteristics (PRTE(x)) but also on the relative intensity of the intervention between units with different covariate values ($w^{\text{IE}}_{\text{PRTE}}(x)$). For example, a tuition subsidy for higher education (e.g., through tax credits) may induce more high-school graduates to attend college. If X denotes the student's family income, then the interventional effect associated with the tuition subsidy is governed by both the interventional effects among students with the same family incomes (PRTE(x)) and the relative responsiveness to the tuition subsidy between students from different income backgrounds

$(w^{\text{IE}}_{\text{PRTE}}(x))$.

2.2 Marginal Interventional Effects

Now consider a class of treatment interventions that can be indexed by a nonnegative scalar δ: $\{\mathcal{I}_\delta : \delta \in [0, M] \text{ where } M > 0\}$. Denote by $O = (X, A, Y)$ the observed data under the status quo and by $O_\delta = (X_\delta, A_\delta, Y_\delta)$ the data that would be observed under intervention $\mathcal{I}_\delta$. Throughout the paper, we maintain the following assumptions.

Assumption 1. *Congruity:* $O_0 = O$.

Assumption 2. *Continuity:* $\lim_{\delta \downarrow 0} \mathbb{E}[A_\delta] = \mathbb{E}[A_0]$.

Assumption 3. *Change:* $\mathbb{E}[A_\delta] - \mathbb{E}[A_0] \neq 0$ *when* $\delta > 0$.

Assumption 4. *No feedback:* $X_\delta = X$ *for all* $\delta \in [0, M]$.

Assumption 1 (*congruity*) means that $\mathcal{I}_0$ can be viewed as a non-intervention that maintains the status quo. The effect of the intervention $\mathcal{I}_\delta$ can therefore be gauged by comparing outcomes under $\mathcal{I}_\delta$ and $\mathcal{I}_0$. Assumption 2 (*continuity*) means that the mean level of treatment is right continuous with respect to δ. In the tuition subsidy example, if δ denotes the amount of the tuition subsidy, then this assumption implies that when the subsidy approaches zero, the proportion of students attending college converges to its baseline level. This assumption is satisfied, for example, in a latent index model that specifies $A_\delta = \mathbb{I}(\delta + g(X) - V \geq 0)$, where $g(\cdot)$ is a function of X, and V denotes an unobserved unit-specific cost of receiving treatment (see Heckman and Vytlacil 2005 for a detailed discussion of latent index models). In this model, we have sure continuity, i.e., $\lim_{\delta \downarrow 0} A_\delta = A$ for all units, which implies continuity in expectation. Sure continuity, however, is not required for Assumption 2 to hold. Consider, for example, a stochastic intervention that sets each unit's treatment status to be a random draw from a Bernoulli distribution with probability $\pi_\delta(X)$. In this case, we no longer have $\lim_{\delta \downarrow 0} A_\delta = A$, because A_δ is random for all units with $\pi_\delta(x) \in (0, 1)$, no matter how small δ is. But Assumption 2 will still hold provided that $\lim_{\delta \to 0} \pi_\delta(x) = \pi(x)$ for all $x \in \text{supp}(X)$.

Assumption 3 (*change*) stipulates that when $\delta > 0$, the intervention $\mathcal{I}_\delta$ induces a change in the proportion of treated units. This assumption does not rule out the possibility that the intervention

induces some units into treatment ($A_\delta - A_0 = 1$) and some other units out of treatment ($A_\delta - A_0 = -1$). If we additionally assume that $A_\delta \geq A_0$ (or $A_\delta \leq A_0$) for all units, such as in the aforementioned latent index model, $\mathbb{E}[A_\delta] - \mathbb{E}[A_0]$ can be interpreted as the proportion of units induced into treatment under intervention $\mathcal{I}_\delta$. Finally, Assumption 4 (*no feedback*) states that all interventions under consideration may induce changes only to the treatment and posttreatment variables.

For a class of interventions satisfying Assumptions 1-4, we define the marginal interventional effect (MIE) as the limit of IE as δ approaches zero, assuming that it exists. That is,

$$\text{MIE} = \lim_{\delta \downarrow 0} \frac{\mathbb{E}[Y_\delta] - \mathbb{E}[Y_0]}{\mathbb{E}[A_\delta] - \mathbb{E}[A_0]}. \tag{3}$$

Thus, MIE reflects the per capita effect of an infinitesimal intervention in the class $\{\mathcal{I}_\delta : \delta \in [0, M]\}$. Following Carneiro *et al.* (2010), we use the word "marginal" to highlight the infinitesimal nature of the intervention being considered. The MIE is useful because it allows us to evaluate the relative impact of different types of interventions at the margin. For example, the MIE under a uniform tuition subsidy applied to all students may be different from that under a means-tested financial aid program.

The MIE is similar to Carneiro *et al.*'s MPRTE, which is defined as the limit of PRTE(x):

$$\text{MPRTE}(x) = \lim_{\delta \downarrow 0} \frac{\mathbb{E}[Y_\delta | X = x] - \mathbb{E}[Y_0 | X = x]}{\pi_\delta(x) - \pi_0(x)},$$

where $\pi_0(x) \stackrel{\Delta}{=} \mathbb{E}[A_0 | X = x]$ and $\pi_\delta(x) \stackrel{\Delta}{=} \mathbb{E}[A_\delta | X = x]$ are the propensity scores of treatment given $X = x$ before and after intervention $\mathcal{I}_\delta$. Under suitable regularity conditions that allow us to exchange limits and integration and apply L'Hôpital's rule, we can deduce from Equation (2) that

$$\begin{aligned} \text{MIE} &= \mathbb{E}\big[\lim_{\delta \downarrow 0} \frac{(\pi_\delta(X) - \pi_0(X))}{\mathbb{E}[(\pi_\delta(X) - \pi_0(X))]} \text{MPRTE}(X) \big] \\ &= \mathbb{E}\big[\underbrace{\frac{\dot{\pi}_0(X)}{\mathbb{E}[\dot{\pi}_0(X)]}}_{\stackrel{\Delta}{=} w^{\text{MIE}}_{\text{MPRTE}}(X)} \text{MPRTE}(X) \big], \end{aligned} \tag{4}$$

where $\dot{\pi}_0(x) \stackrel{\Delta}{=} \partial \pi_\delta(x) / \partial \delta |_{\delta = 0}$. Thus MIE is a weighted mean of MPRTE(X), where the weights

are proportional to the derivative of the propensity score at the status quo. These weights reflect the relative intensity of an infinitesimal intervention among units with covariate values x. For example, $w^{\text{MIE}}_{\text{MPRTE}}(x)$ will likely be greater for low-income students under a means-tested financial aid program than under a uniform tuition subsidy.

In an influential study of the "marginal returns" to college, Carneiro *et al.* (2011) estimated MPRTE(x) under several stylized policy interventions, such as additive and proportional changes in everyone's propensity score of attending college (given both background characteristics X and a set of instrumental variables such as distance to college; see Section 4), and evaluated the "overall" MPRTEs by averaging the corresponding MPRTE(x)'s over the marginal distribution of X, i.e., $\mathbb{E}[\text{MPRTE}(X)]$. This approach implicitly assumes that $w^{\text{MIE}}_{\text{MPRTE}}(x) = 1$, ruling out the possibility that an intervention may differ in strength between individuals with different background characteristics. Real-world interventions, however, are often "preferential" or "targeted." A college outreach program, for instance, will likely induce a larger increase in college attendance among low-income youth than among high-income youth. The effects of such interventions are therefore better captured or approximated by MIE than by the unweighted mean of MPRTE(X).

3 Identification and Estimation under Unconfoundedness

In this section, we discuss the identification and estimation of IE and MIE under the assumption of unconfoundedness, i.e., no unobserved confounding of the treatment-outcome relationship given a set of observed pretreatment covariates X. This assumption may hold by design in a stratified randomized experiment or is maintained by the researcher in an observational study. We focus on interventions that satisfy Assumptions 1-4. In addition, we assume that $\dot{\pi}_0(x) = \partial \pi_\delta(x)/\partial\delta|_{\delta=0}$ exists for all $x \in \text{supp}(X)$ and that it is either a known function of x or a known function of the baseline propensity score $\pi(x)$ (which may be unknown). These conditions are satisfied, for example, if $\pi_\delta(x)$ is a smooth function of δ and $\pi(x)$. Because $\pi_0(x) = \pi(x)$ (by Assumption 1), we henceforth use $\pi_0(x)$ to denote the baseline propensity score.

3.1 General Identification Results

Let $Y_\delta(a)$ denote the potential outcome associated with treatment status a under intervention $\mathcal{I}_\delta$. To identify IE and MIE, we invoke the following assumptions for all $\delta \in [0, M]$.

Assumption 5. *System invariance: For any* $a \in \{0, 1\}$, $Y_\delta(a) = Y_0(a)$.

Assumption 6. *Consistency: For any* $a \in \{0, 1\}$, $Y_\delta = Y_\delta(a)$ *if* $A_\delta = a$.

Assumption 7. *Unconfoundedness: For any* $a \in \{0, 1\}$, $Y_\delta(a) \perp\!\!\!\perp A_\delta | X$.

Assumption 8. *Support:* $\{x : \pi_\delta(x) - \pi_0(x) \neq 0\} \subset \{x : 0 < \pi_0(x) < 1\}$

Assumption 5 (*system invariance*) states that the potential outcomes $Y(0)$ and $Y(1)$ are unaffected by the intervention. In other words, the intervention is not allowed to change the outcome other than through changing treatment status. This assumption is violated if the intervention affects a unit's potential outcome either "directly" (i.e., via pathways other than changing A) or via other units' treatment status (i.e., interference). For example, when analyzing the interventional effect associated with a college outreach program on earnings, Assumption 1 will be violated if an increase in college attendance rate leads to a more competitive labor market among college-goers and a less competitive labor market among non-college-goers, shifting everyone's potential earnings (i.e., a general-equilibrium effect; see Heckman *et al.* 1998). Assumptions 6 (*consistency*) states that under intervention $\mathcal{I}_\delta$, a unit's observed outcome equals its potential outcome under the observed treatment status. This assumption implies that for a given unit, $Y_\delta(a)$ is fixed, thus also ruling out the possibility of interference. Assumption 7 (*unconfoundedness*) means that among units with the same covariate values, treatment assignment is as-if random, i.e., independent of potential outcomes. Finally, assumption 8 (*support*) requires the increment of the propensity score $\pi_\delta(x) - \pi_0(x)$ to be zero at all covariate values where treatment under no intervention is deterministic. This assumption is implied by, but does not imply, the positivity assumption commonly invoked for identifying ATE, which states $0 < \pi_0(x) < 1$ for all $x \in \text{supp}(X)$.

Under Assumptions 5-7, for every $x \in \{x : \pi_\delta(x) - \pi_0(x) \neq 0\}$, we have

$$\text{PRTE}_\delta(x) = \frac{\mathbb{E}[Y_\delta - Y_0 | X = x]}{\mathbb{E}[A_\delta - A_0 | X = x]}$$

$$
\begin{aligned}
&= \frac{\mathbb{E}[Y_\delta(0) + A_\delta(Y_\delta(1) - Y_\delta(0)) - Y_0(0) - A_0(Y_0(1) - Y_0(0))|X = x]}{\mathbb{E}[A_\delta - A_0|X = x]} && \text{(consistency)} \\
&= \frac{\mathbb{E}[Y_0(0) + A_\delta(Y_0(1) - Y_0(0)) - Y_0(0) - A_0(Y_0(1) - Y_0(0))|X = x]}{\mathbb{E}[A_\delta - A_0|X = x]} && \text{(system invariance)} \\
&= \frac{\mathbb{E}[(A_\delta - A_0)(Y_0(1) - Y_0(0))|X = x]}{\mathbb{E}[A_\delta - A_0|X = x]} \\
&= \frac{\mathbb{E}[(A_\delta - A_0)|X = x]\mathbb{E}[Y_0(1) - Y_0(0)|X = x]}{\mathbb{E}[A_\delta - A_0|X = x]} && \text{(unconfoundedness)} \\
&= \text{CATE}(x). && (5)
\end{aligned}
$$

Thus, by Equation (2), the IE associated with intervention $\mathcal{I}_\delta$ is a weighted mean of CATE(X):

$$
\text{IE}_\delta = \mathbb{E}\big[\underbrace{\big(\frac{(\pi_\delta(X) - \pi_0(X))}{\mathbb{E}[(\pi_\delta(X) - \pi_0(X))]}\big)}_{\triangleq w^{\text{IE}_\delta}_{\text{CATE}}(X)} \text{CATE}(X)\big]. \tag{6}
$$

The CATE, in turn, is identified as the conditional mean difference given X, i.e., $\mathbb{E}[Y|X, A = 1] - \mathbb{E}[Y|X, A = 0]$. From Equation (6), we can see that IE_δ is identified as long as CATE(x) is identified over $\{x : \pi_\delta(x) - \pi_0(x) \neq 0\}$. It does not require CATE(x) to be identified over the full support of X, unlike the ATE.

Given Equation (4) and the fact that $\text{MPRTE}(x) = \lim_{\delta \downarrow 0} \text{PRTE}_\delta(x) = \text{CATE}(x)$, MIE is also a weighted mean of CATE(X)

$$
\text{MIE} = \mathbb{E}\big[\underbrace{\frac{\dot{\pi}_0(X)}{\mathbb{E}[\dot{\pi}_0(X)]}}_{\triangleq w^{\text{MIE}}_{\text{CATE}}(X)} \text{CATE}(X)\big]. \tag{7}
$$

Thus, for a given class of interventions, the MIE weight $w^{\text{MIE}}_{\text{CATE}}(x)$ is proportional to $\dot{\pi}_0(x)$, which gauges the relative strength of an infinitesimal intervention at covariate values x.

3.2 Special Cases

From Equations (6) and (7), we can see that IE_δ and MIE depend on $\mathcal{I}_\delta$ only through the propensity score $\pi_\delta(X)$ and its local derivative with respect to δ. In other words, different classes of interventions with the same form of $\pi_\delta(X)$ are equivalent in their IE_δ and MIE. Thus we may focus on classes of interventions that are equivalent up to $\pi_\delta(X)$. Below, we consider a few special cases

where $\pi_\delta(X)$ depends on X only through the baseline propensity score $\pi_0(X)$.

First, let us consider three stylized interventions akin to those proposed in Zhou and Xie (2020): (a) $\pi_\delta(x) = \min\{1, \pi_0(x) + \delta\}$; (b) $\pi_\delta(x) = \min\{1, \pi_0(x)e^\delta\}$; and (c) $\pi_\delta(x) = \min\{1, 1 - (1 - \pi_0(x))e^{-\delta}\}$. Put in words, intervention (a) is "neutral" in the sense that it boosts everyone's propensity score by the same amount (until it reaches one). In this case, it is clear that $w^{\text{MIE}}_{\text{CATE}}(x) \propto \dot{\pi}_0(x) = 1$ if $\pi_0(x) < 1$; otherwise $w^{\text{MIE}}_{\text{CATE}}(x) = 0$. Hence, the corresponding MIE is an unweighted mean of CATE(X) over $\{x : 0 \leq \pi_0(x) < 1\}$, which equals ATE if $\text{supp}(\pi_0(X)) \subset [0, 1)$. Intervention (b) multiplies everyone's propensity score by the same factor (until it reaches one). It is "disequalizing" in the sense that those who are more likely to be treated under the status quo tend to receive a higher increment to their propensity scores. Under this intervention, $w^{\text{MIE}}_{\text{CATE}}(x) \propto \dot{\pi}_0(x) = \pi_0(x)$ if $\pi_0(x) < 1$; otherwise $w^{\text{MIE}}_{\text{CATE}}(x) = 0$. Thus the corresponding MIE is a weighted mean of CATE(X) over $\{x : 0 \leq \pi_0(x) < 1\}$, where the weight is proportional to the baseline propensity score. If $\text{supp}(\pi_0(X)) \subset [0, 1)$, this estimand is equal to ATT. Contrary to intervention (b), intervention (c) is "equalizing" because $\dot{\pi}_0(x) = 1 - \pi_0(x)$, meaning that those who are less likely to be treated under the status quo receive a higher increment to their propensity scores. In this case, MIE is also a weighted mean of CATE(X), where the weight is proportional to $1 - \pi_0(x)$. This estimand, not surprisingly, is equal to the average treatment effect for the untreated (ATU). These stylized interventions are summarized in the first three rows of Table 1. Note that MIE is identified if and only if $\{x : \dot{\pi}_0(x) \neq 0\} \subset \{x : 0 < \pi_0(x) < 1\}$, which implies that intervention (b) is always identified, whereas interventions (a) and (c) are identified only when the baseline propensity score is strictly positive for everyone ($\text{supp}(\pi_0(X)) \subset (0, 1]$).

Now consider the "incremental propensity score intervention" (IPSI) proposed in Kennedy (2019) (the last row of Table 1):

$$\pi_\delta(x) = \frac{e^\delta \pi_0(x)}{1 - \pi_0(x) + e^\delta \pi_0(x)}. \tag{8}$$

Here, we reparameterize the δ parameter in Equation (1) of Kennedy (2019) as e^δ so that $\mathcal{I}_0$ corresponds to a non-intervention. The IPSI is interesting in multiple aspects. First, Equation (8) implies that $\pi_\delta(x) \neq \pi_0(x)$ only when $\pi_0(x) \in (0, 1)$. Thus, the corresponding IE and MIE are identified even if the positivity assumption does not hold, because units whose baseline propensity

Table 1: Four stylized interventions and the associated marginal interventional effects.

$\pi_\delta(x)$	MIE weight ($\propto \dot{\pi}_0(x)$)	Connection to existing estimands	Positivity required for identification?
$\min\{1, \pi_0(x) + \delta\}$	$\propto \mathbb{I}(\pi_0(x) < 1)$	ATE if $\text{supp}(\pi_0(X)) \subset [0, 1)$	Partially ($\text{supp}(\pi_0(X)) \subset (0, 1]$)
$\min\{1, \pi_0(x)e^\delta\}$	$\propto \pi_0(x)\mathbb{I}(\pi_0(x) < 1)$	ATT if $\text{supp}(\pi_0(X)) \subset [0, 1)$	No
$\min\{1, 1 - (1 - \pi_0(x))e^{-\delta}\}$	$\propto 1 - \pi_0(x)$	ATU	Partially ($\text{supp}(\pi_0(X)) \subset (0, 1]$)
$\frac{e^\delta \pi_0(x)}{1-\pi_0(x)+e^\delta \pi_0(x)}$	$\propto \pi_0(x)(1 - \pi_0(x))$	ATO	No

Note: ATE = average treatment effect; ATT = average treatment effect for the treated; ATU = average treatment effect for the untreated; ATO = average treatment for the overlap population.

scores are zero or one would not be affected by the intervention. In this regard, it is similar to intervention (b). Second, as Kennedy noted, e^δ represents an odds ratio, "indicating how the intervention changes the odds of receiving treatment" (pp. 646-47):

$$e^\delta = \frac{\pi_\delta(x)/\big(1 - \pi_\delta(x)\big)}{\pi_0(x)/\big(1 - \pi_0(x)\big)}.$$

Hence, the IPSI multiplies everyone's odds of receiving treatment by the same constant (e^δ); in other words, it increases everyone's *log-odds* of receiving treatment by the same constant (δ). Thus, $\pi_\delta(x)$ can be viewed as the propensity score under a latent index model of treatment where $A_\delta = \mathbb{I}(\delta + g(X) - V \geq 0)$ and V follows a standard logistic distribution. Note that this model implies an intervention under which each unit's treatment status A_δ will be fixed given her observed pretreatment covariates X and unobserved variable V. This interpretation is different from that of Kennedy (2019), who interprets the IPSI as a *stochastic* intervention that sets each unit's treatment status to be an independent random draw from a Bernoulli distribution with probability $\pi_\delta(X)$. In contexts where an individual's treatment status cannot be directly manipulated (to follow a random draw), the former interpretation of the IPSI may be more realistic than the stochastic intervention perspective.

Third, under the IPSI, $\dot{\pi}_0(x) = \pi_0(x)(1 - \pi_0(x))$, implying that the increment to the propensity score is largest for units whose baseline propensity scores are around 0.5. The corresponding MIE

becomes

$$\text{MIE}_{\text{IPSI}} = \frac{\mathbb{E}[\pi_0(X)(1-\pi_0(X))\text{CATE}(X)]}{\mathbb{E}[\pi_0(X)(1-\pi_0(X))]}. \tag{9}$$

This estimand coincides with the average treatment effect of overlap (ATO) proposed by Li *et al.* (2018). These authors formulate the ATO as the estimand associated with overlap weights, an alternative to inverse probability weighting in which "each unit's weight is proportional to the probability of that unit being assigned to the opposite group" (p. 390). Moreover, they show that when the propensity score is estimated via a logistic regression model, the overlap weights lead to exact balance between treated and untreated units in the means of all pretreatment covariates. The ATO is akin to the concept of optimally weighted average treatment effect (OWATE; Crump *et al.* 2006). Specifically, Crump *et al.* (2006) consider a class of weighted sample average treatment effects in the form of $\tau_{S,w} = \sum_i w(X_i)\text{CATE}(X_i))/\sum_i w(X_i)$ and show $w^*(X) = \pi(X)(1-\pi(X))$ to be the weight that minimizes $\tau_{S,w}$'s nonparametric variance bound under homoskedasticity (i.e., $\text{Var}[Y|X, A] = \text{constant}$).

Finally, Equation (9) is also often known as the "regression estimand" (Angrist and Pischke 2008), which, under the assumption that $\mathbb{E}[A|X]$ is linear in X, is the probability limit of the coefficient on A in a multiple regression of Y on A and X. While previous authors highlighted the "nonrepresentative nature" of the regression estimand (e.g., Aronow and Samii 2016), we have seen that the regression estimand is not void of substance. Although it is nonrepresentative of the full population, it is *representative of a policy-relevant subpopulation* that would be induced into treatment under the IPSI, an intervention that preserves the relative odds of receiving treatment between different units.

3.3 Connection with Modified Treatment Policies

As noted earlier, under our identification assumptions, interventions with the same form of $\pi_\delta(X)$ are equivalent in their IE_δ and MIE, which motivated our use of $\pi_\delta(X)$ to characterize different types of interventions. Nonetheless, even interventions with the same $\pi_\delta(X)$ can be understood and operationalized in different ways. For example, we have seen that the IPSI can be interpreted as either a deterministic intervention within a latent index model or a stochastic intervention that sets each unit's treatment status to be an independent random draw from Bernoulli($\pi_\delta(X)$). In either

case, the treatment assignment mechanism under $\mathcal{I}_\delta$ is assumed to depend solely on pretreatment characteristics ((X, V) in the latent index model or X for the stochastic intervention). In epidemiology, however, scholars have proposed an alternative type of interventions called modified treatment policies (MTPs; e.g., Robins *et al.* 2004; Taubman *et al.* 2009; Haneuse and Rotnitzky 2013; Young *et al.* 2014), under which treatment status depends, either deterministically or stochastically, on the "natural" value of treatment that would have been observed without the intervention, i.e., A_0. Haneuse and Rotnitzky (2013), for example, considered the impact of a deterministic MTP that shortens the operating time by a modest amount, i.e., $A_\delta = A_0 - \delta$, on post-operation outcomes of lung cancer patients. The daily exercise intervention envisioned by Taubman *et al.* (2009) is also a deterministic MTP (defined as $A^* = \max(30, A)$), although it is not indexed by a scalar.

At first glance, MTPs differ from the interventions we have discussed in that they depend on the "natural" treatment status A_0. However, our definition and identification of IE and MIE does not rule out the possibility that A_δ may depend on A_0. That is, our identification results for IE and MIE should apply to MTPs provided that Assumptions 1-8 hold. In particular, we note that if system invariance holds for an MTP and if unconfoundedness holds under the status quo (i.e., $Y_0(a) \perp\!\!\!\perp A_0|X$), then unconfoundedness also holds under the MTP. This is because under an MTP, $A_\delta|X, A_0$ is either fixed or an independent random draw, implying $A_\delta \perp\!\!\!\perp Y_0(a)|X, A_0$. The latter conditional independence, when combined with $Y_0(a) \perp\!\!\!\perp A_0|X$, implies $A_\delta \perp\!\!\!\perp Y_0(a)|X$, hence $A_\delta \perp\!\!\!\perp Y_\delta(a)|X$ (due to system invariance). Thus, the IE and MIE under an MTP can also be identified using equations (6) and (7), where the interventional propensity score $\pi_\delta(x)$ can be obtained by marginalizing over A_0:

$$\pi_\delta(x) = \pi_0(x) \Pr[A_\delta = 1|x, A = 1] + \big(1 - \pi_0(x)\big) \Pr[A_\delta = 1|x, A = 0].$$

This result echoes Young *et al.* (2014), who show that the expected outcome under an MTP can be identified using the same g-formula that one would normally use for a stochastic intervention that does not depend on A_0, provided that A_0 does not affect the outcome other than through A_δ. In our framework, the latter condition is ensured by system invariance (Assumption 5).

To gain more intuition as to why MTPs can be subsumed under our framework, consider a stylized MTP that keeps all treated units treated but randomly induces untreated units into

treatment with probability δ. That is, $\Pr[A_\delta = 1|x, A = 1] = 1$ and $\Pr[A_\delta = 1|x, A = 0] = \delta$, which implies $\pi_\delta(x) = \pi_0(x) + \delta\big(1 - \pi_0(x)\big)$. Thus, for this MTP, we have $\dot{\pi}_0(x) = 1 - \pi_0(x)$. The corresponding MIE is therefore ATU, akin to the third stylized intervention shown in Table 1. By contrast, if untreated units are induced into treatment with a probability proportional to their baseline propensity score, we will have $\Pr[A_\delta = 1|x, A = 0] = \delta\pi_0(x)$, implying $\pi_\delta(x) = \pi_0(x) + \delta\pi_0(x)(1 - \pi_0(x))$. In this case, $\dot{\pi}_0(x) = \pi_0(x)(1 - \pi_0(x))$, leading to MIE_{IPSI}.

3.4 Estimation

For the first three stylized interventions shown in Table 1, estimation of MIE will be straightforward if positivity holds (or is assumed to hold), in which case we can apply existing estimators of ATE, ATT, or ATU, such as regression-imputation (RI; e.g., Hahn 1998), inverse probability weighting (e.g., Hirano *et al.* 2003), and doubly robust methods (Robins and Rotnitzky 1995). When positivity is potentially violated but Assumption 8 holds, the corresponding estimands become ATE, ATT, or ATU conditional on $0 < \pi_0(X) < 1$. Such estimands are non-smooth functionals of the distribution of (X, A, Y), which can complicate estimation and inference. In practice, we could apply existing estimators of ATE, ATT, or ATU to a trimmed sample that excludes regions of the covariate space where positivity is either theoretically or practically violated. For example, we could first fit a propensity score model and then restrict the analytical sample to units whose estimated propensity scores lie in the interval $[\min_{i:A_i=1} \hat{\pi}(X_i), \max_{i:A_i=0} \hat{\pi}(X_i)]$. This is in fact a common practice for addressing potential positivity violations in the estimation of conventional causal parameters (e.g., Dehejia and Wahba 1999; Carneiro *et al.* 2011). In our context, this practice is additionally justified by the fact that when Assumption 8 holds, the MIE weights are zero for units whose propensities scores are 0 or 1. Nonetheless, it must be acknowledged that the true propensity scores are generally unknown, and trimming methods based on estimated propensity scores can lead to bias and inferential challanges (see Crump *et al.* 2006 and Petersen *et al.* 2012 for more discussion).

For the IPSI, MIE is Equation (9), for which Li *et al.* (2018) proposed the following weighting estimator:

$$\widehat{\text{MIE}}_{\text{IPSI}}^{\text{weighting}} = \frac{\sum_i (1 - \hat{\pi}_0(X_i)) A_i Y_i}{\sum_i (1 - \hat{\pi}_0(X_i)) A_i} - \frac{\sum_i \hat{\pi}_0(X_i)(1 - A_i) Y_i}{\sum_i \hat{\pi}_0(X_i)(1 - A_i)}.$$

An attractive property of this estimator, as noted above, is that when the propensity scores $\pi_0(X)$ are estimated via a logistic regression model, exact balance is achieved between treated and untreated units in the means of the covariates. Alternatively, one can use an RI approach described in Crump *et al.* (2006):

$$\widehat{\text{MIE}}_{\text{IPSI}}^{\text{RI}}=\frac{\sum_i \hat{\pi}_0(X_i)(1-\hat{\pi}_0(X_i))(\hat{\mu}_1(X_i)-\hat{\mu}_0(X_i))}{\sum_i \hat{\pi}_0(X_i)(1-\hat{\pi}_0(X_i))},$$

where $\mu_a(X) \triangleq \mathbb{E}[Y|X, A=a]$ for $a=0,1$. For both the weighting and RI estimators of MIE_{IPSI}, the nonparametric bootstrap can be used to estimate standard errors and confidence intervals. For the RI estimator, Crump *et al.* (2006) show that it achieves the nonparametric efficiency bound if both the propensity score and outcome models are fit using kernel estimators and certain technical assumptions hold, in which case analytical standard errors can be calculated.

Besides the weighting and RI estimators, Equation (9) can also be estimated using Robinson's (1988) "partialing-out" estimator :

$$\widehat{\text{MIE}}_{\text{IPSI}}^{\text{Robinson}}=\frac{\sum_i (Y_i-\hat{\mathbb{E}}[Y_i|X_i])(A_i-\hat{\mathbb{E}}[A_i|X_i])}{\sum_i (A_i-\hat{\mathbb{E}}[A_i|X_i])^2}. \tag{10}$$

The partialing-out estimator is proposed in the context of the partially linear regression (PLR) model. Although the PLR assumes a constant treatment effect, the probability limit of the partialing-out estimator is still a well-defined statistical parameter under treatment effect heterogeneity. Specifically, it is equal to

$$\tau^{\text{Robinson}} = \frac{\mathbb{E}[\text{Cov}[Y, A|X]]}{\mathbb{E}[\text{Var}[A|X]]},$$

which, when A is binary, reduces to Equation (9) (Vansteelandt and Dukes 2020). An advantage of $\widehat{\text{MIE}}_{\text{IPSI}}^{\text{Robinson}}$ over the weighting and RI estimators is that its estimating equation (10) is based on the efficient influence function (EIF) of τ^{Robinson} in the nonparametric model, namely, $(A-\mathbb{E}[A|X])\big(Y-\mathbb{E}[Y|X]-\tau^{\text{Robinson}}(A-\mathbb{E}[A|X])\big)/\text{Var}[A|X]$. As a result, $\widehat{\text{MIE}}_{\text{IPSI}}^{\text{Robinson}}$ is "Neyman-orthogonal" (Chernozhukov *et al.* 2018), meaning that first step estimation of the nuisance functions $\hat{\mathbb{E}}[Y|X=x]$ and $\hat{\mathbb{E}}[A|X=x]$ has no first-order effect on the influence function of $\widehat{\text{MIE}}_{\text{IPSI}}^{\text{Robinson}}$. This property suggests that $\widehat{\text{MIE}}_{\text{IPSI}}^{\text{Robinson}}$ may still achieve $\sqrt{n}$-consistency even if the nuisance functions

are estimated using data-adaptive/machine learning methods, facilitating what Chernozhukov et al. call debiased machine learning (DML). In particular, if both of these nuisance function estimates are consistent and converge at faster-than-$n^{-1/4}$ rates, and if cross-fitting is used to render the empirical process term asymptotically negligible, $\widehat{\text{MIE}}^{\text{Robinson}}_{\text{IPSI}}$ will be $\sqrt{n}$-consistent, asymptotically normal, and semiparametric efficient. Its standard errors can thus be estimated through the empirical variance of its estimated EIF.

For all of the four stylized interventions described above, the interventional propensity score $\pi_\delta(X)$ is a known function of the baseline propensity score $\pi_0(X)$; so is its local derivative $\dot{\pi}_0(X) = \partial\pi_\delta(X)/\partial\delta|_{\delta=0}$. In general, if $\dot{\pi}_0(X) = \lambda(\pi_0(X))$ where $\lambda(\cdot)$ is a known function, the MIE (i.e., Equation 7) can be estimated by the following RI estimator:

$$\widehat{\text{MIE}}^{\text{RI}} = \frac{\sum_i \lambda(\hat{\pi}_0(X_i))(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))}{\sum_i \lambda(\hat{\pi}_0(X_i))}. \tag{11}$$

Alternatively, one can use the EIF of the corresponding MIE (see Crump *et al.* 2006) to construct a Neyman-orthogonal estimating equation, where the nuisance functions can be estimated using data-adaptive/machine learning methods. Finally, for the special case where $\dot{\pi}_0(x)$ is known (such as in an experimental setting where $\pi_\delta(x)$ is known), we can either use an RI estimator similar to Equation (11) (where $\lambda(\hat{\pi}_0(X_i))$ is replaced by $\dot{\pi}_0(X_i)$) or construct a Neyman-orthogonal and doubly robust estimating equation based on its EIF (see Hirano *et al.* 2003).

3.5 Illustration

We illustrate the MIE estimators described above using data from a right heart catheterization (RHC) study in the U.S. (Connors *et al.* 1996) that collected the survival outcomes of 5,735 hospitalized adult patients following randomized assignment to RHC. RHC is a diagnostic procedure for directly measuring cardiac function in critically ill patients which can cause clinical complications. Using propensity-score-based matching and multivariate regression modeling, Connors et al. (1996) find that after adjusting for confounding, RHC was associated with increased 30-day mortality.

We employ all pretreatment covariates X potentially relating to the decision to use RHC measured in the study. A full description of these variables is available at `https://biostat.app.vumc.org/wiki/pub/Main/DataSets/rhc.html`. Among the 65 observed covariates (21 continu-

ous, 24 binary, and 20 dummy variables from breaking up 5 categorical covariates), distributions of several key covariates differ substantially between the control and treatment groups (see Hirano and Imbens 2001, Table 2). Our outcome of interest is survival 180 days after potential receipt of the treatment. This dataset has most recently been analyzed in Li *et al.* (2018), who implement a set of weighting estimators for estimating ATE, ATT, and ATO.

In this application, we assume that positivity holds, i.e., $\text{supp}(\pi_0(X)) \subset (0, 1)$, which implies that the MIEs for the four stylized interventions correspond to ATE, ATT, ATU, and ATO and that all of them are nonparametrically identified. For each of these estimands, we employ three estimators. First, we employ a parametric IPW approach, using standard ATE, ATT, and ATU weights for the first three interventions and the ATO weight (Li *et al.* 2018) for the IPSI. In all cases, we estimate the propensity scores using a logistic regression of treatment on all covariates. Next, we employ a parametric RI approach, where the outcome model is estimated using a logistic regression that includes the treatment and all covariates. Note that this specification of the outcome model implies a constant treatment effect on the log-odds scale but not on the probability scale. Finally, we implement a DML procedure for each of the four interventions. Specifically, for the first three interventions, we use the doubly robust estimating equations for ATE, ATT, and ATU described in Chernozhukov *et al.* (2018). For the IPSI, we use Robinson's partialing-out estimator (10). In the DML approach, the outcome and propensity score models are both fit using a super learner (van der Laan *et al.* 2007) composed of the generalized linear model (GLM), Lasso, and random forest, and the final estimates are obtained using five-fold cross-fitting. Standard errors for the IPW and RI estimators are estimated using the nonparametric bootstrap with 1,000 replications; standard errors for the DML estimators are estimated from the sample variances of the estimated EIFs.

Table 2 presents our estimates of MIE for the four stylized interventions, i.e., ATE, ATT, ATU, and ATO. To make our results comparable to those reported in Li *et al.* (2018), we multiply our point estimates and standard errors (which are on the probability scale) by 100. All of our estimators suggest that receiving RHC leads to a higher mortality rate than not applying RHC. We find that the MIE estimates under DML are all somewhat smaller than those obtained from the parametric IPW and RI approaches; moreover, despite the use of machine learning methods to reduce model misspecification bias, the estimated standard errors under DML are comparable to

Table 2: MIE estimates under four stylized interventions for the RHC data.

$\pi_\delta(x)$	parametric IPW	parametric RI	DML
$\min\{1, \pi_0(x) + \delta\}$ (ATE)	-5.43 (1.68)	-5.72 (1.33)	-4.56 (1.20)
$\min\{1, \pi_0(x)e^\delta\}$ (ATT)	-5.62 (1.89)	-5.80 (1.37)	-5.13 (1.30)
$\min\{1, 1-(1-\pi_0(x))e^{-\delta}\}$ (ATU)	-5.32 (2.13)	-5.67 (1.32)	-4.22 (1.33)
$\frac{e^\delta \pi_0(x)}{1-\pi_0(x)+e^\delta \pi_0(x)}$ (ATO)	-5.88 (1.35)	-5.76 (1.32)	-5.15 (1.32)

Note: MIE = marginal interventional effect; RHC = right heart catheterization; IPW = inverse probability weighting; RI = regression-imputation; DML = debiased machine learning. Numbers in parentheses are standard errors.

those under parametric RI and, except for ATO, considerably smaller than those under parametric IPW. Under both parametric IPW and DML, our point estimate for the ATO is higher than that for ATE (in absolute value), suggesting that an IPSI might induce into treatment individuals who were particularly vulnerable to its negative consequences.

4 Identification and Estimation with Instrumental Variables

From Section 3, we have seen that under unconfoundedness, IE and MIE can be identified as weighted means of CATEs. The unconfoundedness assumption, however, is strong, untestable, and unlikely to hold in many applications. In observational studies where unconfoundedness is deemed implausible in light of subject matter knowledge or in randomized trials with treatment noncompliance, researchers often seek to identify causal effects using an instrumental-variable (IV) approach (which, to be sure, entails an alternative set of strong and untestable assumptions). In what follows, we demonstrate that IE and MIE can also be identified with IVs within the framework of marginal treatment effects (MTE; Heckman and Vytlacil 1999, 2005).

4.1 The Generalized Roy Model

The MTE framework builds on the generalized Roy model for discrete choices (Roy 1951; Heckman and Vytlacil 2005). As before, let A denote a binary treatment, $Y(a)$ the potential outcome under treatment status a, and X a vector of pretreatment covariates. Following Zhou and Xie (2019), we

write the outcome equations as

$$Y(0) = \mu_0(X) + \epsilon \tag{12}$$

$$Y(1) = \mu_1(X) + \epsilon + \eta \tag{13}$$

where $\mu_0(X) = \mathbb{E}[Y(0)|X]$, $\mu_1(X) = \mathbb{E}[Y(1)|X]$, the error term ϵ captures all unobserved factors that affect the baseline outcome ($Y(0)$), and the error term η captures all unobserved factors that affect the treatment effect ($Y(1)-Y(0)$). In general, the error terms ϵ and η need not be statistically independent of X, although they have zero conditional means by construction. Under Assumption 6 (*consistency*), the observed outcome Y can be linked to the potential outcomes through the so-called switching regression model (Quandt 1958, 1972):

$$\begin{aligned} Y &= (1 - A)Y(0) + AY(1) \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X))A + \epsilon + \eta A. \end{aligned} \tag{14}$$

Treatment assignment is represented by a latent index model. Let I_A be a latent tendency for treatment, which depends on both observed (Z) and unobserved (V) factors:

$$I_A = \mu_A(Z) - V \tag{15}$$

$$A = \mathbb{I}(I_A \geq 0), \tag{16}$$

where $\mu_A(Z)$ is an unspecified function, V is a latent random variable representing unobserved, individual-specific resistance to treatment, assumed to be continuous with a strictly increasing distribution function. The Z vector includes all components of X, but it also includes some IVs ($Z\backslash X$), i.e., exogenous variables that affect only the treatment status A. Moreover, we assume that at least one component of $Z\backslash X$ is continuous. The key assumptions associated with equations (12)-(16) are

Assumption 9. *Independence:* $(\epsilon, \eta, V) \perp\!\!\!\perp Z|X$.

Assumption 10. *Relevance:* $\mu_A(Z)$ *is a nondegenerate function of* Z *given* X.

The latent index model characterized by equations (15) and (16), combined with Assumptions

9-10, is equivalent to the Imbens-Angrist (1994) assumptions of independence and monotonicity for the interpretation of IV estimands as local average treatment effects (LATE) (Vytlacil 2002). Specifically, the separability of $\mu_A(Z)$ and V in equation (15) implies that a change in Z (say from z_1 to z_2) will induce the latent tendency I_A to change in the same direction for all units, thus ruling out the so-called "defiers." Given Assumptions 9-10, the latent resistance V is allowed to be correlated with ϵ and η in a general way. For example, research considering heterogeneous returns to schooling has argued that individuals may self-select into college on the basis of their anticipated gains. In this case, V will be negatively correlated with η, as individuals with higher values of η tend to have lower levels of unobserved resistance U.

4.2 Marginal Treatment Effects

To define MTE, we rewrite the treatment assignment equations (15) and (16) as

$$\begin{aligned} A &= \mathbb{I}(F_{V|X}(\mu_A(Z)) - F_{V|X}(V) \geq 0) \\ &= \mathbb{I}(\pi(Z) - U \geq 0), \end{aligned} \tag{17}$$

where $F_{V|X}(\cdot)$ is the cumulative distribution function (CDF) of V given X, and $\pi(Z) = \Pr(A = 1|Z) = F_{V|X}(\mu_A(Z))$ denotes the propensity score given Z. By definition, $U = F_{V|X}(V)$ follows a standard uniform distribution. From Equation (17) we can see that Z affects treatment status only through the propensity score $\pi(Z)$. The property that Z affects treatment status only through the propensity score in an additively separable latent index model is called index sufficiency (Heckman and Vytlacil 2005).

The MTE is defined as the expected treatment effect conditional on pretreatment covariates $X = x$ and the "normalized" latent variable $U = u$:

$$\begin{aligned} \text{MTE}(x, u) &= \mathbb{E}[Y(1) - Y(0)|X = x, U = u] \\ &= \mathbb{E}[\mu_1(X) - \mu_0(X) + \eta|X = x, U = u] \\ &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta|X = x, U = u]. \end{aligned} \tag{18}$$

Since U is the CDF of V, the variation of $\text{MTE}(x, u)$ over values of u reflects how treatment effect

varies across different quantiles of the unobserved resistance to treatment (given X). Alternatively, $\text{MTE}(x,u)$ can be interpreted as the average treatment effect among individuals with covariates $X = x$ and the propensity score $\pi(Z) = u$ who are indifferent between treatment or not (i.e., $\pi(Z) = U$). A wide range of causal estimands, such as ATE and ATT, can be expressed as weighted averages of $\text{MTE}(x,u)$. To obtain population-level causal effects, $\text{MTE}(x,u)$ needs to be marginalized twice, first over a distribution of U given X and then over the marginal distribution of X. The weights that link MTE to ATE, ATT, and ATU are given in Heckman *et al.* (2006).

Given Equations (12)-(16) and Assumptions 6, 9, and 10, $\text{MTE}(x,u)$ can be identified using the method of local instrumental variables (LIV). To see how it works, let us consider the conditional mean of the observed outcome Y given $X = x$ and the propensity score $\pi(Z) = p$. According to Equation (14), we have

$$\begin{aligned}
\mathbb{E}[Y|X=x,\pi(Z)=p] &= \mathbb{E}[\mu_0(X) + (\mu_1(X)-\mu_0(X))A + \epsilon + \eta A | X=x, \pi(Z)=p] \\
&= \mu_0(x) + (\mu_1(x)-\mu_0(x))p + \mathbb{E}[\eta|A=1,X=x,\pi(Z)=p]p \\
&= \mu_0(x) + (\mu_1(x)-\mu_0(x))p + \int_0^p \mathbb{E}[\eta|X=x,U=u]du \\
&= \mu_0(x) + \int_0^p \text{MTE}(x,u)du \qquad (19)
\end{aligned}$$

Thus, MTE can be identified as the partial derivative of Equation (19) with respect to p:

$$\text{MTE}(x,p) = \frac{\partial \mathbb{E}[Y|X=x,\pi(Z)=p]}{\partial p} \qquad (20)$$

Since $\mathbb{E}[Y|X=x,\pi(Z)=p]$ is a function of observed data, the above equation means that $\text{MTE}(x,u)$ is identified at all values of u within $\text{supp}(\pi(Z)|X=x)$, the conditional support of $\pi(Z)$ given $X = x$. In other words, $\text{MTE}(x,u)$ is identified over $\text{supp}(X,\pi(Z))$, the support of the joint distribution of X and $\pi(Z)$.

4.3 Identification of IE and MIE

As in Section 3, we focus on interventions that satisfy Assumptions 1-6. We assume that $\dot{\pi}_0(z) = \partial\pi_\delta(z)/\partial\delta|_{\delta=0}$ exists for all $z \in \text{supp}(Z)$ and that it is either a known function of z or a known function of the baseline propensity score $\pi(z)$. Because $\pi_0(Z) = \pi(Z)$ (by Assumption 1), we

henceforth use $\pi_0(Z)$ to denote the baseline propensity score.

Assumptions 5-6 imply that the outcome models (12)-(13) are invariant under interventions. Thus the observed outcome Y_δ under intervention $\mathcal{I}_\delta$ can be written as

$$\begin{aligned} Y_\delta &= (1 - A_\delta)Y(0) + A_\delta Y(1) \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X))A_\delta + \epsilon + \eta A_\delta, \end{aligned} \tag{21}$$

Analogous to Equation (19), we have

$$\begin{aligned} \mathbb{E}[Y_\delta] &= \mathbb{E}\mathbb{E}[\mu_0(X) + (\mu_1(X) - \mu_0(X))A_\delta + \epsilon + \eta A_\delta | Z] \\ &= \mathbb{E}\big[\mu_0(X) + (\mu_1(X) - \mu_0(X))\pi_\delta(Z) + \int_0^{\pi_\delta(Z)} \mathbb{E}[\eta | X, U = u] du\big] \\ &= \mathbb{E}\big[\mu_0(X) + \int_0^{\pi_\delta(Z)} \mathrm{MTE}\big(X, u\big) du\big]. \end{aligned}$$

Substituting the above expression into Equation (1) yields

$$\begin{aligned} \mathrm{IE}_\delta &= \frac{\mathbb{E}[Y_\delta] - \mathbb{E}[Y_0]}{\mathbb{E}[A_\delta] - \mathbb{E}[A_0]} \\ &= \frac{\mathbb{E}\big[\int_{\pi_0(Z)}^{\pi_\delta(Z)} \mathrm{MTE}(X, u) du\big]}{\mathbb{E}\big[\pi_\delta(Z) - \pi_0(Z)\big]}. \end{aligned} \tag{22}$$

Thus, IE_δ is identified if for each x, $\mathrm{MTE}(x, u)$ is identified for all $u \in [\pi_0(Z), \pi_\delta(Z)]$. Because $\mathrm{MTE}(x, u)$ is identified for all $u \in \mathrm{supp}(\pi_0(Z)|X = x)$, a sufficient condition for IE_δ to be identified is therefore $\mathrm{supp}(\pi_\delta(Z)|X = x) \subset \mathrm{supp}(\pi_0(Z)|X = x)$, that is, given each covariate value x, the support of the interventional propensity score $\pi_\delta(Z)$ is contained in the support of the baseline propensity score $\pi_0(Z)$. This condition is more likely to hold if the IVs are strong, i.e., if they induce substantial exogenous variation in the baseline propensity score conditional on X.

Under appropriate regularity conditions that allow us to apply L'Hôpital's rule, the Leibniz integral rule, and exchange differentiation/limits and integration, the corresponding MIE can be written as

$$\mathrm{MIE} = \lim_{\delta \to 0} \frac{\mathbb{E}[Y_\delta] - \mathbb{E}[Y_0]}{\mathbb{E}[A_\delta] - \mathbb{E}[A_0]}$$

$$
\begin{aligned}
&= \lim_{\delta \to 0} \frac{\frac{\partial \mathbb{E}\left[\int_{\pi_0(Z)}^{\pi_\delta(Z)} \mathrm{MTE}(X,u)du\right]}{\partial \delta}}{\frac{\partial \mathbb{E}\left[\pi_\delta(Z)-\pi_0(Z)\right]}{\partial \delta}} \\
&= \lim_{\delta \to 0} \frac{\mathbb{E}\left[\frac{\partial \int_{\pi_0(Z)}^{\pi_\delta(Z)} \mathrm{MTE}(X,u)du}{\partial \delta}\right]}{\mathbb{E}\left[\frac{\partial \left[\pi_\delta(Z)-\pi_0(Z)\right]}{\partial \delta}\right]} \\
&= \lim_{\delta \to 0} \frac{\mathbb{E}\left[\dot{\pi}_\delta(Z)\mathrm{MTE}\big(X, \pi_\delta(Z)\big)\right]}{\mathbb{E}\left[\dot{\pi}_\delta(Z)\right]} \\
&= \mathbb{E}\left[\frac{\dot{\pi}_0(Z)}{\mathbb{E}\left[\dot{\pi}_0(Z)\right]}\mathrm{MTE}\big(X, \pi_0(Z)\big)\right]. \qquad (23)
\end{aligned}
$$

Thus, MIE is a weighted mean of $\mathrm{MTE}\big(X, \pi_0(Z)\big)$. Similar to the identification formula for MIE under unconfoundedness, the weight is proportional to the local derivative of the propensity score, which gauges the relative intensity of an infinitesimal intervention among units with $Z = z$. Yet, the "building block" here is not $\mathrm{CATE}(x)$, but $\mathrm{MTE}\big(x, \pi_0(z)\big)$, the marginal treatment effect evaluated at x and $u = \pi_0(z)$. A remarkable fact about Equation (23) is that unlike conventional causal estimands such as ATE, MIE is a weighted mean of $\mathrm{MTE}\big(X, \pi_0(Z)\big)$, not of $\mathrm{MTE}(X, U)$. It means that support conditions are not required to identify MIE, because $\mathrm{MTE}\big(x, \pi_0(z)\big)$ is identified by $\partial \mathbb{E}[Y|x, \pi_0(z)]/\partial \pi_0(z)$ over the entire support of $(X, \pi_0(Z))$.

Similar to the setting of unconfoundedness, we can envision stylized interventions where $\dot{\pi}_0(z)$ is a known function of z or a known function of the baseline propensity score $\pi_0(z)$. For example, in a "neutral" intervention where $\pi_\delta(z) = \min\{1, \pi_0(z) + \delta\}$, $\dot{\pi}_0(z)$ is equal to $\mathbb{I}(\pi_0(z) < 1)$, and the corresponding MIE will be an unweighted mean of $\mathrm{MTE}\big(X, \pi_0(Z)\big)$ over $\{Z : \pi_0(Z) < 1\}$. Yet, unlike the setting of unconfoundedness, the MIE under such a neutral intervention does not reduce to ATE. Instead, it reflects the average treatment effect among units who are at the margin of receiving treatment, i.e., $\mathbb{E}[Y(1) - Y(0)|U = \pi_0(Z)]$. We can also consider disequalizing or equalizing interventions similar to those discussed in Section 3.2, under which those more likely to be treated under the status quo receive a higher or lower increment to their propensity scores. For example, if a college outreach program targets low-income and racial minority neighborhoods, it may induce into college more "unlikely college-goers." Such a program would be better approximated by an equalizing intervention where $\dot{\pi}_0(z)$ is a decreasing function of $\pi_0(z)$.

Finally, if there is no unobserved selection, we have $\big(Y(1), Y(0)\big) \perp\!\!\!\perp A|X$, which implies $(\epsilon, \eta) \perp\!\!\!\perp V|X$. In this case, it is easy to show that $\mathrm{MTE}(x, u) = \mathrm{CATE}(x)$ for any u. Thus, Equation

(23) becomes

$$\begin{aligned}\text{MIE} &= \frac{\mathbb{E}\big[\dot{\pi}_0(Z)\text{CATE}\big(X\big)\big]}{\mathbb{E}\big[\dot{\pi}_0(Z)\big]} \\ &= \frac{\mathbb{E}\big[\mathbb{E}\big[\dot{\pi}_0(Z)|X\big]\text{CATE}\big(X\big)\big]}{\mathbb{E}\mathbb{E}\big[\dot{\pi}_0(Z)|X\big]} \\ &= \mathbb{E}\big[\frac{\dot{\pi}_0(X)}{\mathbb{E}[\dot{\pi}_0(X)]}\text{CATE}(X)\big],\end{aligned}$$

where the last equality is due to the fact that $\mathbb{E}\big[\dot{\pi}_0(Z)|X\big] = \mathbb{E}[\partial\pi_\delta(Z)/\partial\delta|_{\delta=0}|X] = \partial_\delta\mathbb{E}[\pi_\delta(Z)|X]/\partial\delta|_{\delta=0} = \partial_\delta\pi_\delta(X)/\partial\delta|_{\delta=0} = \dot{\pi}_0(X)$. Thus, in the absence of unobserved selection, the identification formula for MIE reduces to that derived under unconfoundedness, which does not involve any IV. A similar result can be obtained for IE.

4.4 Estimation

Since we assume that $\dot{\pi}_0(z)$ is either a known function of z or a known function of the baseline propensity score $\pi_0(z)$, the weights in Equation (23) are either known or can be estimated through a propensity score model. To simplify our exposition, we henceforth assume that $\dot{\pi}_0(Z)$ is known, noting that it can be replaced by an estimator of it when unknown. We outline two approaches to estimating MIE: a plug-in approach based on Equation (23) and a doubly robust approach for settings where $\dot{\pi}_0(z)$ is a known function of x and $\pi_0(z)$.

First, Equation (23) suggests a plug-in estimator of MIE:

$$\widehat{\text{MIE}}^{\text{plug-in}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{\pi}_0(Z_i)}{\sum_{i=1}^{n}\dot{\pi}_0(Z_i)/n}\widehat{\text{MTE}}\big(X_i, \pi_0(Z_i)\big), \tag{24}$$

where $\widehat{\text{MTE}}\big(x, \pi_0(z)\big)$ can be evaluated via Equation (20), i.e., by taking the partial derivative of $\mathbb{E}[Y|x, \pi_0(z)]$ with respect to $\pi_0(z)$. In practice, it can be difficult to estimate $\mathbb{E}[Y|x, \pi_0(z)]$ and its partial derivative nonparametrically, especially when X is high-dimensional. Thus, empirical work using MTE often makes two simplifying assumptions (e.g., Carneiro and Lee 2009; Carneiro *et al.* 2011; Maestas *et al.* 2013). First, it is typically assumed that (X, Z) is jointly independent

of (ϵ, η, V). Under this assumption, $\text{MTE}(x, u)$ is additively separable in x and u:

$$\begin{aligned}\text{MTE}(x, u) &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta | X = x, U = u] \\ &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta | U = u]. \end{aligned} \tag{25}$$

In addition, the conditional means of $Y(0)$ and $Y(1)$ given X are often specified as linear in parameters: $\mu_0(X) = \beta_0^T X$ and $\mu_1(X) = \beta_1^T X$. Given the linear specification and the additive separability of MTE, $\mathbb{E}[Y|X = x, \pi_0(Z) = p]$ can be written as

$$\mathbb{E}[Y|X = x, \pi_0(Z) = p] = \beta_0^T x + (\beta_1 - \beta_0)^T x p + \underbrace{\int_0^p \mathbb{E}[\eta | U = u] du}_{\triangleq K(p)}, \tag{26}$$

where $K(p)$ is an unknown function of p that can be estimated either parametrically or nonparametrically.

In the special case where the error terms (ϵ, η, V) are assumed to be jointly normal with zero means and an unknown covariance matrix Σ, the generalized Roy model characterized by Equations (12), (13), (15), and (16) is fully parameterized, and the unknown parameters $(\beta_1, \beta_0, \gamma, \Sigma)$ can be estimated via maximum likelihood. This model specification has a long history in econometrics and is usually referred to as the "normal switching regression model" (Heckman 1978; see Winship and Mare 1992 for a review). With the joint normality assumption, Equation (25) reduces to

$$\text{MTE}(x, u) = (\beta_1 - \beta_0)^T x + \frac{\sigma_{\eta V}}{\sigma_V} \Phi^{-1}(u) \tag{27}$$

where $\sigma_{\eta V}$ is the covariance between η and V, σ_V is the standard deviation of V, and $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function. By substituting the maximum likelihood estimates of $(\beta_1, \beta_0, \sigma_{\eta V}, \sigma_V)$ into Equation (27), we can obtain a parametric estimate of $\text{MTE}(x, u)$ for any combination of x and u.

The joint normality of error terms is a strong and restrictive assumption. When the errors (ϵ, η, V) are not normally distributed, imposition of normality can lead to substantial bias in the estimates of target parameters (Arabmazar and Schmidt 1982). To avoid this problem, Heckman *et al.* (2006) propose that we fit Equation (26) semiparametrically using Robinson's (1988)

partialing-out procedure. In this case, the estimation of $\text{MTE}(x, u)$ can be summarized in four steps:

1. Estimate the propensity scores using a standard logit/probit model, and denote them as $\hat{\pi}(Z)$;
2. Fit local linear regressions of Y, X, and $X\hat{\pi}(Z)$ on $\hat{\pi}(Z)$ and extract their residuals e_Y, e_X, and $e_{X\hat{\pi}(Z)}$;
3. Fit a simple linear regression of e_Y on e_X and $e_{X\hat{\pi}(Z)}$ (with no intercept) to estimate the parametric component of Equation (26), i.e., $(\beta_0, \beta_1 - \beta_0)$, and store the residuals of Y from this regression as $e_Y^* = Y - \hat{\beta}_0^T X - (\hat{\beta}_1 - \hat{\beta}_0)^T X\hat{\pi}(Z)$.
4. Fit a local quadratic regression (Fan and Gijbels 1996) of e_Y^* on $\hat{\pi}(Z)$ to estimate $K(p)$ and its derivative $K'(p)$.

A semiparametric estimator of MTE is then given by

$$\widehat{\text{MTE}}(x, u) = (\hat{\beta}_1 - \hat{\beta}_0)^T x + \hat{K}'(u). \tag{28}$$

Either the parametric or the semiparametric estimator of $\widehat{\text{MTE}}(x, u)$ can be used to construct a plug-in estimate of MIE (Equation 24). Both approaches, however, rely on correct specification of the outcome model $\mathbb{E}[Y|X = x, \pi(Z) = p]$. We now outline a third approach that is more robust to potential misspecification of the outcome model. Specifically, in settings where $\dot{\pi}_0(z)$ is a known function of x and $\pi_0(z)$, we can construct a doubly robust estimator of MIE based on the EIF of weighted average derivatives (Powell *et al.* 1989; Newey and Stoker 1993). Define $w(X, \pi_0(Z)) \triangleq \dot{\pi}_0(Z)/\mathbb{E}[\dot{\pi}_0(Z)]$, $m(X, \pi_0(Z)) \triangleq \mathbb{E}[Y|X, \pi(Z)]$, and denote by $f(X, \pi_0(Z))$ the density of $(X, \pi_0(Z))$. The MIE can then be written as

$$\text{MIE} = \mathbb{E}[w(X, \pi_0(Z)) \frac{\partial m(X, \pi_0(Z))}{\partial \pi_0(Z)}]. \tag{29}$$

To facilitate nonparametric estimation of such a functional, Newey and Stoker (1993) invoke the assumption that $w(X, \pi_0(Z))f(X, \pi_0(Z)) = 0$ at the boundary of the support of $\pi_0(Z)$. This assumption is reasonable if the positivity assumption holds such that no units have a propensity score of zero or one, or if $w(X, 0) = w(X, 1) = 0$ by construction, such as in the case where

$\dot{\pi}_0(z) = \pi_0(z)(1-\pi_0(z))$. Under this assumption and the assumption that $\pi_0(Z)$ is known, the EIF of Equation (29) is

$$\psi(x, \pi_0(z), y) = w(X, \pi_0(Z))\frac{\partial m(X, \pi_0(Z))}{\partial \pi_0(Z)} + l(X, \pi_0(Z))[Y - m(X, \pi_0(Z))] - \text{MIE}, \tag{30}$$

where

$$l(X, \pi_0(Z)) = -\frac{\partial w(X, \pi_0(Z))}{\partial \pi_0(Z)} - w(X, \pi_0(Z))\frac{\partial \log f(\pi_0(Z)|X)}{\partial \pi_0(Z)}. \tag{31}$$

Equation (30) suggests an estimating equation for MIE. Suppose we have fit the following models: $\hat{\pi}_0(z)$ for $\pi_0(z)$, $\hat{m}(x, \pi_0(z))$ for $m(x, \pi_0(z))$, and $\hat{f}(\pi_0(z)|x)$ for the conditional density $f(\pi_0(z)|x)$. An MIE estimator can then be constructed as

$$\widehat{\text{MIE}}^{\text{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left\{w(X_i, \hat{\pi}_0(Z_i))\frac{\partial \hat{m}(X_i, \hat{\pi}_0(Z_i))}{\partial \hat{\pi}_0(Z_i)} + \hat{l}(X_i, \hat{\pi}_0(Z_i))[Y_i - m(X_i, \hat{\pi}_0(Z_i))]\right\}, \tag{32}$$

where $\hat{l}(X_i, \hat{\pi}_0(Z_i))$ is obtained by substituting $\hat{f}(\hat{\pi}_0(Z_i)|X_i)$ into Equation (31). This estimator is doubly robust: if the propensity score model $\pi_0(z)$ is correctly specified, then $\widehat{\text{MIE}}^{\text{DR}}$ will be consistent if either $m(X_i, \pi_0(Z))$ and $\partial m(X, \pi_0(Z))/\partial \pi_0(Z)$ are consistently estimated or $\partial \log f(\pi_0(Z)|X)/\partial \pi_0(Z)$ is consistently estimated (Chernozhukov *et al.* 2016; Rothe and Firpo 2019). In practice, we can estimate $m(X_i, \pi_0(Z))$ and $\partial m(X, \pi_0(Z))/\partial \pi_0(Z)$ using either the parametric or the semiparametric estimator described above. For the partial derivative $\partial \log f(\pi_0(Z)|X)/\partial \pi_0(Z)$, we can invoke a location shift model for $\pi_0(Z)|X$:

$$\pi_0(Z) = \mathbb{E}[\pi_0(Z)|X] + \epsilon, \quad \epsilon \perp\!\!\!\perp X. \tag{33}$$

For this model, we can estimate $\mathbb{E}[\pi_0(Z)|X]$ using either a GLM or data-adaptive methods such as Lasso. We can then estimate $\partial \log f(\pi_0(Z)|X)/\partial \pi_0(Z)$ using a kernel estimator for the density of ϵ and its derivative:

$$\frac{\partial \widehat{\log f(\pi_0(Z_i)|X_i)}}{\partial \pi_0(Z_i)} = \frac{\hat{\phi}'(\hat{\epsilon}_i)}{\hat{\phi}(\hat{\epsilon}_i)},$$

where $\phi(\cdot)$ denotes the density function of ϵ. If we assume $\epsilon \sim N(0, \sigma_\epsilon^2)$, the above expression reduces to $-\hat{\epsilon}_i/\widehat{\sigma_\epsilon^2}$.

Since both the (parametric or semiparametric) plug-in and the doubly robust estimators of MIE

rely on estimated propensity scores $\hat{\pi}_0(Z_i)$, their inference is not straightforward. To account for the estimation uncertainty of $\hat{\pi}_0(Z_i)$, it is best to use the nonparametric bootstrap to estimate the standard errors and confidence intervals of $\widehat{\text{MIE}}^{\text{plug-in}}$ and $\widehat{\text{MIE}}^{\text{DR}}$.

4.5 Connections to Existing Work

The PRTE defined by Heckman and Vytlacil (2001) and the MPRTE defined by Carneiro *et al.* (2010) are similar to IE and MIE except that they are conditional on $X = x$. Analogous to Equations (22) and (23), these quantities can be written as weighted means of $\text{MTE}(x, U)$ and $\text{MTE}(x, \pi_0(Z))$, respectively:

$$\text{PRTE}_\delta(x) = \frac{\mathbb{E}\big[\int_{\pi_0(Z)}^{\pi_\delta(Z)} \text{MTE}\big(X, u\big) du | X = x\big]}{\mathbb{E}\big[\pi_\delta(Z) - \pi_0(Z) | X = x\big]},$$
$$\text{MPRTE}(x) = \mathbb{E}\Big[\frac{\dot{\pi}_0(Z)}{\mathbb{E}\big[\dot{\pi}_0(Z) | X = x\big]} \text{MTE}\big(X, \pi_0(Z)\big) | X = x\Big].$$

To evaluate the overall impact of a marginal policy change, these authors consider the average of $\text{MPRTE}(x)$ over the marginal distribution of X, i.e., $\mathbb{E}[\text{MPRTE}(X)]$ (Carneiro *et al.* 2010, 2011). As noted in Section 2, this estimand can only characterize interventions that vary in intensity between units different values of $Z \backslash X$ (the IVs) but not between units with different values of X. This is because the weights that link $\text{MTE}\big(X, \pi_0(Z)\big)$ to $\mathbb{E}[\text{MPRTE}(X)]$ have a mean of one regardless of the covariate values x: $\mathbb{E}\big[\dot{\pi}_0(Z) / \mathbb{E}[\dot{\pi}_0(Z) | X = x] \big| X = x\big] = 1$. This restriction is unrealistic and undesirable because real-world interventions often target individuals and communities with particular background characteristics, such as race and socioeconomic status, which typically serve as pretreatment covariates rather than IVs in empirical analyses (see Zhou and Xie 2020 for a detailed discussion).

In contrast to $\mathbb{E}[\text{MPRTE}(X)]$, MIE is a weighted mean of $\text{MPRTE}(X)$ (Equation 4), where the weights $\dot{\pi}_0(X) / \mathbb{E}[\dot{\pi}_0(X)]$ capture the relative intensity of an infinitesimal intervention between units with different values of the pretreatment covariates. From this perspective, $\text{MPRTE}(x)$ could be used as an intermediate quantity for us to evaluate MIE. Nonetheless, estimation of $\text{MPRTE}(x)$ is not straightforward, as the weights that link $\text{MTE}\big(x, \pi_0(Z)\big)$ to $\text{MPRTE}(x)$ involve the conditional density of $\big(\pi_0(Z), \dot{\pi}_0(Z)\big)$ given $X = x$. Since X is often high-dimensional, estimation of

these weights is practically challenging and often tackled via ad hoc methods (e.g., Carneiro *et al.* 2011). By contrast, the plug-in and doubly robust estimators that we introduced in Section 4.4 use Equation (23) directly and do not hinge on a correct model for $f(\pi_0(Z), \dot{\pi}_0(Z)|X = x)$.

More recently, Zhou and Xie (2019, 2020) proposed a modified approach to defining and estimating MPRTE. This approach is based on a redefinition of MTE as the expected treatment effect conditional on the baseline propensity score $\pi_0(Z)$, rather than the whole vector of pretreatment covariates X, as well as the latent resistance to treatment U:

$$\widetilde{\text{MTE}}(p, u) \triangleq \mathbb{E}[Y(1) - Y(0)|\pi_0(Z) = p, U = u].$$

$\widetilde{\text{MTE}}(p, u)$ is a bivariate function that summarizes how treatment effects vary by observed ($\pi_0(Z)$) and unobserved (U) characteristics. Because a unit is treated if and only if $\pi_0(Z) \geq U$, Zhou and Xie propose a new version of MPRTE:

$$\widetilde{\text{MPRTE}}(p) \triangleq \widetilde{\text{MTE}}(p, p).$$

$\widetilde{\text{MPRTE}}(p)$ is a univariate function that reveals how treatment effects vary by the propensity score $\pi_0(Z)$ among units who are at the margin of treatment, i.e., units for whom $\pi_0(Z) = U$. These authors then consider policy changes of the form $\pi_\delta(Z) = \pi_0(Z) + \delta\lambda(\pi_0(Z))$, where $\lambda(\cdot)$ is a known function, and show the unconditional MPRTE to be a weighted average of $\widetilde{\text{MPRTE}}(p)$:

$$\widetilde{\text{MPRTE}} = \frac{\mathbb{E}\big[\lambda(\pi_0(Z))\widetilde{\text{MPRTE}}(\pi_0(Z))\big]}{\mathbb{E}\big[\lambda(\pi_0(Z))\big]}. \tag{34}$$

Compared with Carneiro et al.'s approach, the above approach to studying policy effects is "unconditional" in that it allows a policy change to vary in intensity between individuals with different baseline propensity scores $\pi_0(Z)$, regardless of whether the variation in $\pi_0(Z)$ stems from the pretreatment covariates X or the IVs ($Z \backslash X$). Yet, it is also somewhat restrictive because it accommodates only policy changes where the increment $\pi_\delta(Z) - \pi_0(Z)$ is a function of the baseline propensity score. In fact, Equation (34) is a consequence of Equation (23) when $\dot{\pi}_0(Z) = \lambda(\pi_0(Z))$

where $\lambda(\cdot)$ is a known function:

$$\begin{aligned}
\text{MIE} &= \mathbb{E}\big[\frac{\dot{\pi}_0(Z)}{\mathbb{E}\big[\dot{\pi}_0(Z)\big]}\text{MTE}\big(X, \pi_0(Z)\big)\big] \\
&= \frac{\mathbb{E}[\lambda(\pi_0(Z))\text{MTE}\big(X, \pi_0(Z)\big)]}{\mathbb{E}\big[\lambda(\pi_0(Z))\big]} \\
&= \frac{\mathbb{E}\big[\lambda(\pi_0(Z))\mathbb{E}[\text{MTE}\big(X, \pi_0(Z)\big)|\pi_0(Z)]\big]}{\mathbb{E}\big[\lambda(\pi_0(Z))\big]} \\
&= \frac{\mathbb{E}\big[\lambda(\pi_0(Z))\widetilde{\text{MPRTE}}(\pi_0(Z))\big]}{\mathbb{E}\big[\lambda(\pi_0(Z))\big]}.
\end{aligned}$$

The last equality is due to the fact that $\mathbb{E}[\text{MTE}\big(X, \pi_0(Z)\big)|\pi_0(Z)] = \widetilde{\text{MTE}}\big(\pi_0(Z), \pi_0(Z)\big)$ (see Zhou and Xie 2019). Thus, our identification formula (23) can be viewed a generalization of Zhou and Xie's approach.

4.6 Illustration

We illustrate the MIE estimators described in Section 4.4 by reanalyzing data from Carneiro *et al.*'s (2011) study on the economic returns to college (see also Zhou and Xie 2020). The dataset consists of 1,747 white males at ages 16-22 in 1979, drawn from the National Longitudinal Survey of Youth (NLSY) 1979. Treatment (A) is college attendance defined by having attained any post-secondary education by 1991. The treated group consists of 865 individuals, and the untreated group consists of 882 individuals. The outcome Y is the natural logarithm of hourly wage around 1991, defined as the average of deflated (to 1983 dollars) non-missing hourly wages reported between 1989 and 1993. The pretreatment variables (X) include linear and quadratic terms of mother's years of schooling, number of siblings, the Armed Forces Qualification Test (AFQT) score adjusted by years of schooling, permanent local log earnings at age 17 (county log earnings averaged between 1973 and 2000), permanent local unemployment rate at age 17 (state unemployment rate averaged between 1973 and 2000), as well as a dummy variable indicating urban residence at age 14 and cohort dummies. The instrumental variables ($Z \backslash X$) include: (a) the presence of a four-year college in the county of residence at age 14; (b) local wage in the county of residence at age 17; (c) local unemployment rate in the state of residence at age 17; and (d) average tuition in public four-year colleges in the county of residence at age 17; and (e) their interactions with mother's years of

schooling, number of siblings, and the adjusted AFQT score. Following the original study, we also include four variables in X but not in Z: years of experience in 1991, years of experience in 1991 squared, local log earnings in 1991, and local unemployment rate in 1991. More details about the data can be found in the online appendix of Carneiro *et al.* (2011).

We consider four stylized interventions akin to those introduced in Section 3.2, except that the baseline and interventional propensity scores now are the conditional probabilities of treatment given Z rather than X. Thus, the local derivatives of the interventional propensity scores are: (a) $\dot{\pi}_0(Z) = 1\mathbb{I}\big(\pi_0(Z) < 1\big)$; (b) $\dot{\pi}_0(Z) = \pi_0(Z)\mathbb{I}\big(\pi_0(Z) < 1\big)$; (c) $\dot{\pi}_0(Z) = 1 - \pi_0(Z)$; (d) $\dot{\pi}_0(Z) = \pi_0(Z)(1 - \pi_0(Z))$. In particular, intervention (b) is disequalizing in that it nudges into college more students with relatively high baseline propensity scores. Conversely, intervention (c) is equalizing, inducing more "unlikely college-goers" into college. Following Carneiro *et al.* (2011), we discard 67 observations whose estimated propensity scores lie outside the interval $[\min_{i:A_i=1} \hat{\pi}(Z_i), \max_{i:A_i=0} \hat{\pi}(Z_i)]$ (which is $[0.0324, 0.9775]$). For the restricted sample, we assume that positivity holds ($0 < \pi_0(Z) < 1$), such that the indicator functions that appear in $\dot{\pi}_0(Z)$ in cases (a) and (b) can be ignored. For each of the four interventions, we estimate MIE using three estimators: a parametric plug-in estimator where MTE is estimated using the normal switching regression model (i.e. Equation 27), a semiparametric plug-in estimator where MTE is estimated using the partialing-out procedure described in Section 4.4 (i.e., Equation 28), and the doubly robust estimator (32). For the doubly robust estimator, we assume a location-shift model for $\pi_0(Z)|X$ with normal errors, where $\mathbb{E}[\pi_0(Z)|X]$ is estimated using a generalized linear model with a logistic link. Standard errors for these estimators are estimated using the nonparametric bootstrap with 1,000 replications.

Table 3 presents our estimates of MIE for the four stylized interventions. We can see that all four interventions imply substantial marginal effects of college, although point estimates differ among the three estimators. Under the "neutral" intervention, for example, the semiparametric plug-in estimate of MIE is 0.379, suggesting that attending college would translate into a 46.1% ($e^{0.379} - 1 = 0.461$) increase in hourly wages among the affected students. Moreover, the estimated MIE depends considerably on the form of the policy change, especially under the semiparametric plug-in and doubly robust methods. In particular, the equalizing intervention ($\pi_\delta(z) = \min\{1, 1 - (1-\pi_0(z))e^{-\delta}\}$) appears to yield a higher MIE than the other three interventions. This result echoes

Table 3: MIE estimates under four stylized interventions for the NLSY data.

$\pi_\delta(z)$	parametric plug-in	semiparametric plug-in	doubly robust
$\min\{1, \pi_0(z) + \delta\}$	0.318 (0.154)	0.379 (0.169)	0.480 (0.178)
$\min\{1, \pi_0(z)e^{\delta}\}$	0.323 (0.176)	0.265 (0.19)	0.274 (0.224)
$\min\{1, 1 - (1 - \pi_0(z))e^{-\delta}\}$	0.312 (0.148)	0.488 (0.185)	0.677 (0.202)
$\frac{e^{\delta}\pi_0(z)}{1-\pi_0(x)+e^{\delta}\pi_0(z)}$	0.291 (0.150)	0.344 (0.154)	0.471 (0.180)

Note: In the doubly robust estimator, $m(X_i, \pi_0(Z))$ and $\partial m(X, \pi_0(Z))/\partial \pi_0(Z)$ are estimated using the semiparametric method described in Section 4.4, and the partial derivative $\partial \log f(\pi_0(Z)|X)/\partial \pi_0(Z)$ is estimated using a location shift model for $\pi_0(Z)|X$ with normal errors.

Zhou and Xie's (2020) estimates of $\widetilde{\text{MPRTE}}(p)$, which suggest that among marginal entrants (i.e., for whom $\pi(Z) = U$), students who benefit the most from college are located at the low end of the baseline propensity score.

5 Concluding Remarks

In this article, we have expounded the concepts of IE and MIE, which are defined as the per capita effect of a treatment intervention on an outcome of interest and its limit when the size of the intervention approaches zero. Compared with conventional causal estimands, IE and MIE map more closely onto real-world policy changes, which typically "nudge" into (or out of) treatment a small segment of the population who are at or near the margin of participation. These concepts can be seen as the unconditional counterparts of the PRTE and MPRTE parameters proposed in the economics literature (Heckman and Vytlacil 2001, 2005; Carneiro *et al.* 2010, 2011). Yet, in contrast to the context in which PRTE and MPRTE were introduced, IE and MIE are defined without reference to a latent index model and can be identified either under the assumption of unconfoundedness or through the use of IVs.

The generality of IE and MIE allows us to bridge the econometrics literature on policy-relevant causal effects and the scholarship on interventional effects that has independently developed in statistics and epidemiology. We have shown that under unconfoundedness, both IE and MIE can be identified as a weighted mean of CATEs, and, in particular, the MIE associated with the IPSI proposed in Kennedy (2019) coincides with the ATO parameter, which has hitherto been advocated

for purely statistical reasons (Li *et al.* 2018). Without unconfoundedness, we can identify IE and MIE using IVs under Heckman and Vytlacil's latent index model. Specifically, in this framework, MIE can be expressed as a weighted mean of $\text{MTE}(X, \pi_0(Z))$, i.e., the average treatment effect among individuals with pretreatment covariates X who are at the margin of treatment ($U = \pi_0(Z)$). Because $\text{MTE}\big(x, \pi_0(z)\big)$ is identified over the entire support of $(X, \pi_0(Z))$, MIE can be identified even if the variation of $\pi_0(Z)$ given X is very limited. Moreover, we have discussed several different estimation strategies for MIE, including plug-in estimators based on parametric or semiparametric estimates of $\text{MTE}(x, u)$ and a doubly robust approach based on the EIF of weighted average derivatives.

In this article, we restricted our attention to settings with a binary treatment. A growing body of research has considered interventional effects for continuous treatments under the assumption of unconfoundedness (e.g., Díaz and van Der Laan 2012; Díaz and van der Laan 2013; Haneuse and Rotnitzky 2013; Young *et al.* 2014; Rothenhäusler and Yu 2019; Hines *et al.* 2021). For instance, Díaz and van Der Laan (2012) consider interventions that shift the conditional mean of a continuous treatment given pretreatment covariates, and Rothenhäusler and Yu (2019) define "incremental causal effect" as the limit of one particular type of Díaz and van Der Laan's intervention, one in which the shift in treatment is constant across units. Future research could incorporate the concepts of IE and MIE into this line of inquiry, consider a broader array of MIEs than the incremental causal effect defined in Rothenhäusler and Yu (2019), and develop identification and estimation results under alternative assumptions.

Acknowledgements

The authors thank Ang Yu and two reviewers from the Alexander and Diviya Magaro Peer Pre-Review Program for helpful comments.

References

Angrist, J. D. and Pischke, J.-S. (2008) *Mostly Harmless Econometrics*. Princeton university press.

Arabmazar, A. and Schmidt, P. (1982) An investigation of the robustness of the tobit estimator to non-normality. *Econometrica*, **50**, 1055–1063.

Aronow, P. M. and Samii, C. (2016) Does regression produce representative estimates of causal effects? *American Journal of Political Science*, **60**, 250–267.

Carneiro, P., Heckman, J. J. and Vytlacil, E. J. (2010) Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin. *Econometrica*, **78**, 377–394.

Carneiro, P., Heckman, J. J. and Vytlacil, E. J. (2011) Estimating Marginal Returns to Education. *American Economic Review*, **101**, 2754–2781.

Carneiro, P. and Lee, S. (2009) Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, **149**, 191–208.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21**, C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2016) Locally robust eemiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M. *et al.* (1996) The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*, **276**, 889–897.

Crump, R., Hotz, V. J., Imbens, G. and Mitnik, O. (2006) Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand.

Dehejia, R. H. and Wahba, S. (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, **94**, 1053–1062.

Díaz, I. and van Der Laan, M. (2012) Population intervention causal effects based on stochastic interventions. *Biometrics*, **68**, 541–549.

Díaz, I. and van der Laan, M. J. (2013) Assessing the causal effect of policies: an example using stochastic interventions. *The International Journal of Biostatistics*, **9**, 161–174.

Eberhardt, F. and Scheines, R. (2007) Interventions and causal inference. *Philosophy of Science*, **74**, 981–995.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*, vol. 66. London: Chapman and Hall.

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

Haneuse, S. and Rotnitzky, A. (2013) Estimation of the effect of interventions that modify the received treatment. *Statistics in Medicine*, **32**, 5260–5277.

Heckman, J. J. (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrica*, **46**, 931–959.

Heckman, J. J., Lochner, L., Taber, C. *et al.* (1998) General-equilibrium treatment effects: A study of tuition policy. *American Economic Review*, **88**, 381–386.

Heckman, J. J., Urzua, S. and Vytlacil, E. J. (2006) Understanding Instrumental Variables in Models with Essential Heterogeneity. *The Review of Economics and Statistics*, **88**, 389–432.

Heckman, J. J. and Vytlacil, E. J. (1999) Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4730–4734.

Heckman, J. J. and Vytlacil, E. J. (2001) Policy-Relevant Treatment Effects. *American Economic Review*, **91**, 107–111.

Heckman, J. J. and Vytlacil, E. J. (2005) Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, **73**, 669–738.

Hines, O., Diaz-Ordaz, K. and Vansteelandt, S. (2021) Parameterising the effect of a continuous exposure using average derivative effects. *arXiv preprint arXiv:2109.13124*.

Hirano, K. and Imbens, G. W. (2001) Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, **2**, 259–278.

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

Imbens, G. W. and Angrist, J. D. (1994) Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62**, 467–475.

Kennedy, E. H. (2019) Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, **114**, 645–656.

Korb, K. B., Hope, L. R., Nicholson, A. E. and Axnick, K. (2004) Varieties of causal intervention. In *Pacific Rim International Conference on Artificial Intelligence*, 322–331. Springer.

Li, F., Morgan, K. L. and Zaslavsky, A. M. (2018) Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, **113**, 390–400.

Maestas, N., Mullen, K. J. and Strand, A. (2013) Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *The American Economic Review*, **103**, 1797–1829.

Moore, K. L., Neugebauer, R., Van der Laan, M. J. and Tager, I. B. (2012) Causal inference in epidemiological studies with strong confounding. *Statistics in Medicine*, **31**, 1380–1404.

Murphy, S. A., van der Laan, M. J., Robins, J. M. and Group, C. P. P. R. (2001) Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, **96**, 1410–1423.

Naimi, A. I., Rudolph, J. E., Kennedy, E. H., Cartus, A., Kirkpatrick, S. I., Haas, D. M., Simhan, H. and Bodnar, L. M. (2020) Incremental propensity score effects for time-fixed exposures. *Epidemiology*, **32**, 202–208.

Newey, W. K. and Stoker, T. M. (1993) Efficiency of weighted average derivative estimators and index models. *Econometrica: Journal of the Econometric Society*, 1199–1223.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. and Van Der Laan, M. J. (2012) Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, **21**, 31–54.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989) Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 1403–1430.

Quandt, R. E. (1958) The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, **53**, 873–880.

Quandt, R. E. (1972) A New Approach to Estimating Switching Regressions. *Journal of the American Statistical Association*, **67**, 306–310.

Robins, J. M., Hernán, M. A. and Siebert, U. (2004) Effects of multiple interventions. *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, **1**, 2191–2230.

Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.

Robinson, P. M. (1988) Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rothe, C. and Firpo, S. (2019) Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory*, **35**, 1048–1087.

Rothenhäusler, D. and Yu, B. (2019) Incremental causal effects. *arXiv preprint arXiv:1907.13258*.

Roy, A. D. (1951) Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*, **3**, 135–146.

Shpitser, I. and Pearl, J. (2006) Identification of joint interventional distributions in recursive semi-markovian causal models. In *21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI-06/IAAI-06*, 1219–1226.

Taubman, S. L., Robins, J. M., Mittleman, M. A. and Hernán, M. A. (2009) Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology*, **38**, 1599–1611.

Tian, J. (2012) Identifying dynamic sequential plans. *arXiv preprint arXiv:1206.3292*.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, **6**.

Vansteelandt, S. and Dukes, O. (2020) Assumption-lean inference for generalised linear model parameters. *arXiv preprint arXiv:2006.08402*.

Vytlacil, E. (2002) Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, **70**, 331–341.

Winship, C. and Mare, R. D. (1992) Models for Sample Selection Bias. *Annual Review of Sociology*, **18**, 327–50.

Xie, Y. (2013) Population Heterogeneity and Causal Inference. *Proceedings of the National Academy of Sciences*, **110**, 6262–6268.

Young, J. G., Hernán, M. A. and Robins, J. M. (2014) Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic Methods*, **3**, 1–19.

Zhou, X. and Xie, Y. (2019) Marginal treatment effects from a propensity score perspective. *Journal of Political Economy*, **127**, 3070–3084.

Zhou, X. and Xie, Y. (2020) Heterogeneous treatment effects in the presence of self-selection: a propensity score perspective. *Sociological Methodology*, **50**, 350–385.