Received: 3 April 2021 | Accepted: 16 March 2022

DOI: 10.1111/rssb.12538

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS Series B Statistical Methodology B

Multiply robust estimation of causal effects under principal ignorability

Zhichao Jiang[1] | **Shu Yang**[2] | **Peng Ding**[3]

[1]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts, USA

[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

[3]University of California, Berkeley, Berkeley, California, USA

Correspondence
Peng Ding, University of California, Berkeley, CA 94720, USA.
Email: pengdingpku@berkeley.edu

Abstract

Causal inference concerns not only the average effect of the treatment on the outcome but also the underlying mechanism through an intermediate variable of interest. Principal stratification characterizes such a mechanism by targeting subgroup causal effects within principal strata, which are defined by the joint potential values of an intermediate variable. Due to the fundamental problem of causal inference, principal strata are inherently latent, rendering it challenging to identify and estimate subgroup effects within them. A line of research leverages the principal ignorability assumption that the latent principal strata are mean independent of the potential outcomes conditioning on the observed covariates. Under principal ignorability, we derive various nonparametric identification formulas for causal effects within principal strata in observational studies, which motivate estimators relying on the correct specifications of different parts of the observed-data distribution. Appropriately combining these estimators yields triply robust estimators for the causal effects within principal strata. These triply robust estimators are consistent if two of the treatment, intermediate variable and outcome models are correctly specified, and moreover, they are locally efficient if all three models are correctly specified. We show that these estimators arise naturally from either the efficient influence functions in the semiparametric theory or the model-assisted estimators in the survey sampling theory. We evaluate different estimators based on their finite-sample performance through simulation and apply them to two observational studies.

KEYWORDS
noncompliance, principal stratification, sensitivity analysis, surrogate endpoint, truncation by death

1 | INTRODUCTION

Researchers are often interested in understanding the underlying causal mechanism from the treatment to the outcome when an intermediate variable is present between them. This requires proper adjustment for the intermediate variable—naively conditioning on its observed value does not have a valid causal interpretation unless it is essentially randomized conditional on the treatment and covariates (Rosenbaum, 1984). Frangakis and Rubin (2002) propose to estimate causal effects within principal strata, which are defined by the joint potential values of the intermediate variable under both treatment and control. Principal strata act as pretreatment covariates, so the causal effects within them, often referred to as principal causal effects (PCEs), are conceptually the same as the standard subgroup causal effects. PCEs are widely used in applied statistics to deal with noncompliance (Angrist et al., 1996; Frumento et al., 2012; Mealli & Pacini, 2013), truncation by death (Ding et al., 2011; Rubin, 2006; Wang et al., 2017), missing data (Frangakis & Rubin, 1999; Mattei et al., 2014), mediation (Elliott et al., 2010; Gallop et al., 2009; Mattei & Mealli, 2011; Rubin, 2004) and surrogate evaluation (Frangakis & Rubin, 2002; Gilbert & Hudgens, 2008; Huang & Gilbert, 2011; Jiang et al., 2016; Li et al., 2010).

Due to the fundamental problem of causal inference, the two potential values of the intermediate variable are not simultaneously observable, rendering it challenging to identify and estimate PCEs without additional assumptions. Angrist et al. (1996) establish the nonparametric identification of one PCE, often called the complier average causal effect or the local average treatment effect, under the monotonicity and the exclusion restriction (ER). The monotonicity assumes that the treatment changes the intermediate variable only in one direction for any unit, and the ER assumes that the treatment affects the outcome only through the intermediate variable. The identification result of Angrist et al. (1996) has motivated various estimation methods and efficiency theories (Abadie, 2003; Frölich, 2007; Ogburn et al., 2015; Tan, 2006). Although the ER is a standard assumption, it is not plausible when the treatment affects the outcome through pathways other than the intermediate variable. Hirano et al. (2000) give an example of the violation of the ER in a randomized experiment with noncompliance. Moreover, in mediation, truncation by death, and principal surrogate evaluation problems, testing the ER is a scientific question of interest. Thus, we cannot invoke ER *a priori*. Without the ER, Zhang et al. (2008) and Imai (2008) derive the large sample bounds on the PCEs, and Li et al. (2010), Zigler and Belin (2012), and Schwartz et al. (2011) perform model-based Bayesian analyses. Unfortunately, the bounds might be too wide to be informative, whereas the Bayesian analyses could be sensitive to models and priors. Assuming Normal linear outcome models within principal strata, Zhang et al. (2009) and Frumento et al. (2012) estimate the PCEs using the likelihood approach, but these analyses can be sensitive to the modelling assumptions and can be unstable even if the models are correctly specified due to the mixture distributions of the observed data (Feller et al., 2022). Auxiliary covariates or secondary outcomes satisfying additional conditional independence assumptions can help to improve identification and estimation of the PCEs (e.g. Ding et al., 2011; Jiang & Ding, 2021; Jiang et al., 2016; Mattei & Mealli, 2011; Mattei et al., 2013; Mealli & Pacini, 2013; Yang & Small, 2016), but those additional assumptions may be hard to justify without prior knowledge.

We focus on an alternative nonparametric identification strategy under principal ignorability, an assumption in parallel with ignorability for estimating the average causal effect in observational studies (Rosenbaum & Rubin, 1983). Principal ignorability assumes that the observed covariates are adequate for controlling for confounding between the principal strata and outcome. This identification strategy has been popular in applied statistics (Egleston et al., 2009; Feller et al., 2017; Follmann, 2000; Hayden et al., 2005; Hill et al., 2002; Jo & Stuart, 2009; Jo et al., 2011; Stuart & Jo, 2015).

We develop a statistical methodology for estimating the PCEs in both randomized experiments and observational studies under principal ignorability. We first establish three identification formulas for each PCE. These formulas motivate three estimators for each PCE, which rely on correct specifications of two of the following three models:

1. the model of the treatment conditional on the covariates, called the *treatment probability*;
2. the model of the intermediate variable conditional on the treatment and covariates, called the *principal score* with a little abuse of terminology;
3. the model of the mean of the outcome conditional on the treatment, intermediate variable and covariates, called the *outcome mean*.

The existence of multiple estimators for the same parameter hints at the possibility of a combined estimator for each PCE. To guide the construction of principled estimators, we derive the efficient influence functions (EIFs; Bickel et al., 1993) for the PCEs under the nonparametric model. These EIFs motivate novel estimators for PCEs based on the treatment probability, principal score and outcome mean. Interestingly, the novel estimators are triply robust in that they are consistent and asymptotically Normal if any two of the three models in (a)–(c) are correctly specified, and locally efficient if all three models are correctly specified. These results extend the classic doubly robust estimators for the average causal effect in observational studies (Bang & Robins, 2005) and are similar in spirit to the triply robust estimators in other contexts of causal inference (Shi et al., 2020; Tchetgen Tchetgen & Shpitser, 2012; Wang & Tchetgen Tchetgen, 2018). The new triply robust estimators offer additional protection against model misspecification compared to other non-robust estimators. Finally, we establish an equivalence relationship between the triply robust estimation and the model-assisted estimation, extending the existing results on the average causal effect in observational studies (Kang & Schafer, 2007; Little & An, 2004; Lumley et al., 2011; Robins & Rotnitzky, 1998).

Previously, Ding and Lu (2017) establish some preliminary results for estimating the PCEs in randomized experiments including an identification formula based on weighting and the corresponding estimators for each PCE with and without adjusting for covariates. Their model-assisted estimator adjusts for covariates but is neither doubly robust nor semiparametrically efficient. So even in randomized experiments, the theory for estimating the PCEs is incomplete. We discuss a broader class of treatment assignments in both randomized experiments and unconfounded observational studies, providing two additional identification formulas for each PCE and proposing more principled estimators based on the EIFs. Our new estimators outperform those in Ding and Lu (2017) and we recommend using them in data analyses.

The rest of this paper proceeds as follows. Section 2 introduces notation and assumptions for identification. Section 3 presents three different identification formulas and the corresponding estimators of the PCEs. Section 4 derives the EIFs, proposes novel estimators and shows the triple robustness of the estimators. Section 5 uses simulation to evaluate the finite-sample properties of the estimators, and Section 6 applies the novel estimators to two observational studies. Section 7 concludes. The supplementary material contains the technical details including some extensions and the proofs.

2 | NOTATION AND ASSUMPTIONS FOR PRINCIPAL STRATIFICATION

Let $Z_i \in \{0, 1\}$ be the binary treatment, $S_i \in \{0, 1\}$ the binary intermediate variable, Y_i the outcome and X_i a vector of pretreatment covariates for unit $i = 1, \ldots, n$. We adopt the potential outcomes framework under the Stable Unit Treatment Value Assumption, and let S_{iz} and Y_{iz} be the potential values of the intermediate variable and outcome if unit i were to receive treatment condition z $(z = 0, 1)$. The observed intermediate variable and outcome are thus $S_i = Z_i S_{i1} + (1 - Z_i) S_{i0}$ and $Y_i = Z_i Y_{i1} + (1 - Z_i) Y_{i0}$. Assume $\{Z_i, S_{i1}, S_{i0}, Y_{i1}, Y_{i0}, X_i : i = 1, \ldots, n\}$ are independent and identically distributed. Thus, the observed $\{Z_i, S_i, Y_i, X_i : i = 1, \ldots, n\}$ are also independent and identically distributed. For simplicity, we drop the subscript i when no confusion arises.

Frangakis and Rubin (2002) use the joint potential values of the intermediate variable to define the principal stratification variable, $U = (S_1, S_0)$. For a binary intermediate variable, U can be $(0, 0), (1, 0), (0, 1)$, and $(1, 1)$. For the ease of exposition, we will simplify (S_1, S_0) as $S_1 S_0$ throughout the paper. Define the PCE as the average causal effect within a principal stratum:

$$\tau_{s_1 s_0} = \mathbb{E}(Y_1 - Y_0 | U = s_1 s_0), \quad (s_1 s_0 = 00, 10, 11, 01).$$

The scientific meanings of the PCEs vary with the contexts. We review four canonical examples below.

Example 1 (Noncompliance). In noncompliance problems, Z is the treatment assigned, S is the treatment received and Y is the outcome. The principal strata $U = (0, 0), (1, 0), (0, 1), (1, 1)$ are referred to as never-takers, compliers, always-takers and defiers respectively. Angrist and Imbens (1994) and Angrist et al. (1996) propose to estimate τ_{10}, the complier average causal effect, which is also called the local average treatment effect.

Example 2 (Truncation by death). In truncation-by-death problems, Z is the treatment, S is the survival status, and Y is often a measure of the quality of life. Rubin (2006) points out that the only well-defined causal effect is τ_{11}, which characterizes the treatment effect for patients who would survive regardless of the treatment. Other PCEs are not well defined because the quality of life is defined only for survived patients.

Example 3 (Mediation). In mediation analysis, S is the mediator that lies on the causal pathway from the treatment Z to the outcome Y. The subgroup effects τ_{11} and τ_{00} can assess the direct effect of the treatment on the outcome because the treatment does not change the mediator in these two strata (Gallop et al., 2009; Mattei & Mealli, 2011; Rubin, 2004). In contrast, the subgroup effects τ_{10} and τ_{01} are less interpretable because they consist of both direct and indirect effects (VanderWeele, 2011).

Example 4 (Surrogate evaluation). In surrogate evaluation problems, S is the surrogate candidate for the effect of the treatment Z on the outcome Y. Frangakis and Rubin (2002) propose the principal surrogate criterion based on 'causal necessity'. It requires that Z affects Y only if Z affects S, that is, $\tau_{11} = \tau_{00} = 0$. Gilbert and Hudgens (2008) argue that a valid surrogate should also satisfy 'causal sufficiency'. It requires that if the treatment effect on the surrogate is non-zero, then the treatment effect on the outcome is also non-zero, that is, $\tau_{10} \neq 0$ and $\tau_{01} \neq 0$. See Jiang et al. (2016) for a related discussion.

We will focus on the setting with treatment ignorability for both the intermediate variable and outcome, extending the classic treatment ignorability in observational studies.

Assumption 1 (Treatment ignorability). $Z \perp\!\!\!\perp (S_0, S_1, Y_0, Y_1)|X$.

Assumption 1 rules out latent confounding between the treatment and intermediate variable and that between the treatment and outcome. It holds by the design of a randomized experiment, where the treatment is independent of all the potential values and covariates, that is, $Z \perp\!\!\!\perp (S_0, S_1, Y_0, Y_1, X)$; Ding and Lu (2017) focus on this special case. It also holds by the design of a stratified experiment based on a discrete X, where the treatment is independent of all the potential values within each stratum of X. In observational studies, its plausibility relies on whether or not the observed covariates include all the confounders that affect the treatment as well as the outcome and intermediate variable.

Since we do not observe S_1 and S_0 simultaneously, U is not directly observable. As a result, the PCEs are not identifiable without additional assumptions. We impose the standard monotonicity assumption throughout the paper, which helps to identify the distribution of U, even though the individual U_i's are not observed for all units.

Assumption 2 (Monotonicity). $S_1 \geq S_0$.

Assumption 2 requires that the treatment has a non-negative impact on the intermediate variable for all units, which rules out stratum $U = 01$. It holds automatically when $S_0 = 0$, for example, in one-sided noncompliance problems (Sommer & Zeger, 1991) and vaccine trials without immune response under control (Follmann, 2006).

Under Assumptions 1 and 2, two nonparametric identification strategies exist for the PCEs, relying on different additional assumptions. We review them below.

2.1 | Strategy one based on exclusion restriction

The first strategy assumes the ER:

$$\tau_{11} = \tau_{00} = 0. \tag{1}$$

A stronger version of the ER is $Y_1 = Y_0$ for $U = 11$ or 00. Under Assumptions 1, 2, and Equation (1), Angrist and Imbens (1994) and Angrist et al. (1996) establish the nonparametric identification of the complier average causal effect

$$\tau_{10} = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(S|Z=1) - \mathbb{E}(S|Z=0)}.$$

By definition, the PCE τ_{10} represents the effect of the treatment assigned for compliers. Moreover, for compliers with $U = 10$, the treatment assigned is identical to the treatment received, so τ_{10} also measures the effect of the treatment received. This formulation of the noncompliance problem is due to Frangakis and Rubin (2002) where the potential outcome Y_z corresponds to the treatment assigned. It corresponds to the intervention Z in the actual experiment without assuming that S is another hypothetical intervention. However, it might cause notational incoherence with Angrist and Imbens (1994) and Angrist et al. (1996). Angrist and Imbens (1994) index the potential outcome Y_s by the treatment received, and thus enforce the ER assumption automatically; Angrist et al. (1996) index the potential outcome Y_{zs} by both the treatment assigned and received, and reduce it to Y_s under the ER assumption. These different formulations do not cause fundamental differences. An advantage of Frangakis and Rubin (2002)'s formulation is its generality to deal with other problems with intermediate variables. In Example 1 with noncompliance, it allows us to assess the plausibility of the ER by estimating τ_{11} and τ_{00}; see Section 6.1 for more details.

The ER requires that the treatment has no direct effect on the outcome, which is sometimes implausible in open-label randomized experiments. More importantly, it cannot be invoked in problems where τ_{00} and τ_{11} are the quantities of interest, such as truncation by death in Example 2, mediation in Example 3, and surrogate evaluation in Example 4.

Without the ER, the PCEs are not identifiable. Under weak assumptions, the large-sample bounds on the PCEs are often not informative (Imai, 2008; Zhang et al., 2008). In contrast, Bayesian methods often require specifying strong mixture model assumptions and prior distributions (Li et al., 2010; Schwartz et al., 2011; Zigler & Belin, 2012). They are not easy to implement and can be numerically unstable in practice (Feller et al., 2022). Due to these limitations, we focus on another approach assuming principal ignorability.

2.2 | Strategy two based on principal ignorability

The principal ignorability can be viewed as the analogue of the treatment ignorability assumption in unconfounded observational studies.

Assumption 3 (Principal ignorability). $\mathbb{E}(Y_1|U = 11, X) = \mathbb{E}(Y_1|U = 10, X)$ and $\mathbb{E}(Y_0|U = 00, X) = E(Y_0|U = 10, X)$.

Assumption 3 requires that the expectations of the potential outcomes do not vary across principal strata conditional on the covariates. It is widely used in applied statistics (Follmann, 2000; Hill et al., 2002; Jo & Stuart, 2009; Jo et al., 2011; Stuart & Jo, 2015). Under Assumptions 1–2, Assumption 3 is equivalent to

$$\mathbb{E}(Y_1|U = 11, Z = 1, S = 1, X) = \mathbb{E}(Y_1|U = 10, Z = 1, S = 1, X), \tag{2}$$

$$\mathbb{E}(Y_0|U = 00, Z = 0, S = 0, X) = \mathbb{E}(Y_0|U = 10, Z = 0, S = 0, X). \tag{3}$$

The observed stratum ($Z = 1$, $S = 1$) is a mixture of two principal strata $U = 11, 10$. Therefore, Equation (2) means that within the observed stratum ($Z = 1$, $S = 1$), the expectation of the potential outcome Y_1 does not vary across the two principal strata conditional on the covariates. So the conditional expectations in Equation (2) simplify to the observable conditional expectation $\mathbb{E}(Y|Z=1,S=1,X)$. Similarly, Equation (3) means that within the observed stratum ($Z=0$, $S=0$), the two principal strata $U=00$, 10 are ignorable for the expectation of the potential outcome Y_0 conditional on the covariates. So the conditional expectations in Equation (3) simplify to the observable conditional expectation $\mathbb{E}(Y|Z=0,S=0,X)$. Intuitively, principal ignorability simplifies a latent mixture problem to an observed mixture problem. With this assumption, we can treat the subpopulation ($Z=z$, $S=s$) as a mixture of strata defined by the observed covariates, which is easier to deal with than a mixture of latent principal strata.

We start with Assumptions 2 and 3 because they allow for deriving simple identification formulas and easy-to-implement estimators. These estimators are numerically stable and statistically robust. They can be benchmark estimators in data analyses. Nevertheless, their plausibility cannot be validated by the observed data, so they should be made with caution. To supplement the theory under Assumptions 2 and 3, we also propose corresponding sensitivity analysis techniques for the potential violations of these assumptions. Due to the space limit, we include the theoretical results and numerical examples in the supplementary material.

2.3 | Principal ignorability and sequential ignorability in mediation analysis

Before giving the nonparametric identification formulas of the PCEs based on principal ignorability, we comment on its relationship with a commonly used assumption in mediation analysis. We also make a brief comparison of principal stratification and mediation analysis.

A stronger version of Assumption 3 is $Y_z \perp\!\!\!\perp S_{1-z}|(S_z,X)$ for $z=0$, 1. It assumes that conditional on the covariates, the potential outcome depends only on the potential intermediate variable under the same treatment condition, but not the one under a different treatment condition. Importantly, principal ignorability is different from the *sequential ignorability* between the intermediate variable and the outcome, which is a common assumption in mediation analysis (Imai et al., 2010; Pearl, 2001; Tchetgen Tchetgen & Shpitser, 2012). In particular, the sequential ignorability assumes away the dependence between the potential outcome and the potential intermediate variable given the covariates, while principal ignorability allows for such dependence but rules out the dependence between the potential outcome and intermediate variable under different treatment conditions. Hence, the sequential ignorability and principal ignorability focus on different relationships between the potential outcome and the potential intermediate variable and thus do not imply each other. Forastiere et al. (2018) propose a generalized strong principal ignorability and show that under monotonicity, it is equivalent to the sequential ignorability. Their definition does not imply Assumption 3, and thus it is essentially different from the principal ignorability used in the literature.

In general, principal stratification and mediation analysis can be conceptually different. Principal stratification does not require that S is a well-defined intervention as in Examples 2 and 4. In contrast, traditional mediation analysis requires that S is a well-defined intervention on the causal pathway from the treatment to the outcome. VanderWeele (2011) points out this issue, whereas Robins et al. (2020) attempt to relax this assumption with an alternative approach to mediation analysis.

3 | NONPARAMETRIC IDENTIFICATION AND ESTIMATION

3.1 | Identification formulas

To simplify the exposition, define

$$\pi(X) = \mathbb{P}(Z = 1|X), \quad e_u(X) = \mathbb{P}(U = u|X), \quad \mu_{zs}(X) = \mathbb{E}(Y|Z = z, S = s, X)$$

for $u = 10$, 00, 11 and z, $s = 0$, 1. The $\pi(X)$ is the treatment probability given the covariates, also known as the propensity score. The $e_u(X)$ is the principal score which equals the proportion of principal stratum u given the covariates. The $\mu_{zs}(X)$ is the mean of the outcome within the observed group $(Z = z, S = s)$ given the covariates. Let $\pi = \mathbb{E}\{\pi(X)\} = \mathbb{P}(Z = 1)$ and $e_u = \mathbb{E}\{e_u(X)\}$ denote the marginalized treatment probability and principal score over the distribution of the covariates respectively. Thus, π represents the proportion of treated units and e_u represents the proportion of units with $U = u$.

Under Assumption 2, Table 1 shows the relationship between the observed strata defined by (Z, S) and the principal strata. So under Assumptions 1 and 2, the principal scores are identified by

$$e_{10}(X) = p_1(X) - p_0(X), \quad e_{00}(X) = 1 - p_1(X), \quad e_{11}(X) = p_0(X),$$

where $p_z(X) = \mathbb{P}(S = 1|Z = z, X)$ is the probability of the intermediate variable conditional on the treatment and covariates. Analogously, the proportions of principal strata are identified by

$$e_{10} = p_1 - p_0, \quad e_{00} = 1 - p_1, \quad e_{11} = p_0,$$

where $p_z = \mathbb{E}\{p_z(X)\}$ is the marginalized probability of the intermediate variable over the distribution of the covariates. Due to the one-to-one mapping between $\{p_1(X), p_0(X)\}$ and $\{e_{11}(X), e_{00}(X), e_{10}(X)\}$, we call both sets the principal score, and the exact meaning should be clear from the context. The following theorem provides three identification formulas for each PCE.

Theorem 1 (Nonparametric identification). *Suppose that Assumptions 1–3 hold,* $e_u > 0$ *for* $u = 10, 00, 11$, *and* $0 < \pi(x) < 1$ *for all* x *in the support of* X. *The following identification formulas hold for the PCEs.*

(a) *Based on the treatment probability and principal score,*

$$\tau_{10} = \mathbb{E}\left\{\frac{e_{10}(X)}{p_1 - p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{10}(X)}{p_1 - p_0}\frac{1 - S}{1 - p_0(X)}\frac{1 - Z}{1 - \pi(X)}Y\right\},$$

TABLE 1 Principal strata in the observed strata defined by (Z, S) under monotonicity

	$S = 0$	$S = 1$
$Z = 0$	$U \in \{00, 10\}$	$U = 11$
$Z = 1$	$U = 00$	$U \in \{11, 10\}$

$$\tau_{00} = \mathbb{E}\left\{\frac{1-S}{1-p_1}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{00}(X)}{1-p_1}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\},$$
$$\tau_{11} = \mathbb{E}\left\{\frac{e_{11}(X)}{p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{S}{p_0}\frac{1-Z}{1-\pi(X)}Y\right\}.$$

(b) *Based on the treatment probability and outcome mean,*

$$\tau_{10} = \mathbb{E}\left[\frac{SZ/\pi(X) - S(1-Z)/\{1-\pi(X)\}}{p_1 - p_0}\{\mu_{11}(X) - \mu_{00}(X)\}\right],$$
$$\tau_{00} = \mathbb{E}\left[\frac{1-SZ/\pi(X)}{1-p_1}\{\mu_{10}(X) - \mu_{00}(X)\}\right],$$
$$\tau_{11} = \mathbb{E}\left[\frac{S(1-Z)/\{1-\pi(X)\}}{p_0}\{\mu_{11}(X) - \mu_{01}(X)\}\right].$$

(c) *Based on the principal score and outcome mean,*

$$\tau_{10} = \mathbb{E}\left[\frac{p_1(X)-p_0(X)}{p_1-p_0}\{\mu_{11}(X) - \mu_{00}(X)\}\right],$$
$$\tau_{00} = \mathbb{E}\left[\frac{1-p_1(X)}{1-p_1}\{\mu_{10}(X) - \mu_{00}(X)\}\right],$$
$$\tau_{11} = \mathbb{E}\left[\frac{p_0(X)}{p_0}\{\mu_{11}(X) - \mu_{01}(X)\}\right].$$

Theorem 1 gives identification formulas for the PCEs based on three different combinations of the likelihood components. Theorem 1(a) is an extension of Ding and Lu (2017) with an additional weighting term based on the inverse of the treatment probability, which is also mentioned by Jiang and Ding (2021). Theorem 1(b) and (c) are two additional sets of identification formulas.

Below we give some intuition based on only τ_{10} since the discussion for the other two PCEs is similar. Theorem 1(a) expresses τ_{10} as the difference between weighted averages of the outcome under the treatment and control. The weights in the formula consist of two parts: $Z/\pi(X)$ and $(1-Z)/\{1-\pi(X)\}$ correspond to the treatment probability; $e_{10}(X)S/p_1(X)$ and $e_{10}(X)(1-S)/\{1-p_0(X)\}$ correspond to the principal score. Under Assumptions 1 and 2, the conditional expectations of the weights equal

$$\mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}|X\right\} = \mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}|X\right\} = \frac{e_{10}(X)}{e_{10}}, \tag{4}$$

that is, the conditional probability of principal stratum $U = 10$ divided by its unconditional probability.

Theorem 1(b) expresses τ_{10} in terms of the treatment probability and outcome mean. Under principal ignorability, the difference between the outcome means equals

$$\mu_{11}(X) - \mu_{00}(X) = \mathbb{E}(Y_1|U=10,X) - \mathbb{E}(Y_0|U=10,X) = \mathbb{E}(Y_1 - Y_0|U=10,X),$$

which is the PCE for stratum $U = 10$ conditional on X. Under Assumptions 1 and 2, the conditional expectation of the unnormalized weight equals $\mathbb{E}[SZ/\pi(X) - S(1-Z)/\{1-\pi(X)\}|X] =$

$e_{10}(X)$. Dividing this by the normalizing constant, $p_1 - p_0 = e_{10}$, yields $e_{10}(X)/e_{10}$, which, by Bayes' Theorem, is the density ratio of X conditional and unconditional on $U = 10$. Therefore, the identification formula for τ_{10} in Theorem 1(b) averages the conditional PCE over the distribution of X given $U = 10$, which gives the PCE within stratum $U = 10$.

Theorem 1(c) expresses τ_{10} in terms of the principal score and outcome mean. Compared with Theorem 1(b), the treatment probability weighting is replaced with the principal score weighting $\{p_1(X) - p_0(X)\}/(p_1 - p_0)$. Under Assumptions 1 and 2, the weight again equals $e_{10}(X)/e_{10}$. Therefore, similar to the discussion of Theorem 1(b), Theorem 1(c) identifies τ_{10} by averaging the conditional PCE over the distribution of X given $U = 10$.

3.2 | Estimators based on the nonparametric identification formulas

For each PCE, the three identification formulas in Theorem 1 motivate three estimators, which require correct specifications of different parts of the observed-data distribution. For descriptive convenience, we introduce additional notation. Let $\mathbb{P}_n$ denote the empirical average, for example, $\mathbb{P}_n h(V) = n^{-1}\sum_{i=1}^n h(V_i)$ for any $h(V)$. Let $\hat{\pi} = \mathbb{P}_n Z$ be the moment estimator of π. Let $\pi(X; \alpha)$ be a working parametric model for the treatment probability $\pi(X)$, $p_z(X; \gamma)$ a working parametric model for the principal score $p_z(X)$ for $z = 0, 1$, and $\mu_{zs}(X; \beta)$ a working parametric model for the outcome mean $\mu_{zs}(X)$ for $z, s = 0, 1$. Because $e_u(X)$ has a one-to-one mapping to $p_z(X)$, we use $e(X; \gamma)$ to denote a working parametric model for $e_u(X)$. We focus on parametric models here and will consider more flexible estimation strategies later. Based on the maximum likelihood estimation or the method of moments, we obtain estimators $\hat{\alpha}$, $\hat{\gamma}$ and $\hat{\beta}$. Assume they have probability limits α^*, γ^* and β^* respectively. We use $\mathcal{M}$ with subscripts 'tp', 'ps' and 'om' to denote models with the correct specification of the treatment probability, principal score and outcome mean respectively. Therefore, under $\mathcal{M}_{\text{tp}}$, we have $\pi(X; \alpha^*) = \pi(X)$; under $\mathcal{M}_{\text{ps}}$, we have $p_z(X; \gamma^*) = p_z(X)$ and $e_u(X; \gamma^*) = e_u(X)$ for $z = 0, 1$ and $u = 10, 00, 11$; under $\mathcal{M}_{\text{om}}$, we have $\mu_{zs}(X; \beta^*) = \mu_{zs}(X)$ for z, $s = 0, 1$. In addition, we use '+' in the subscript to indicate that more than one model is correctly specified. For example, $\mathcal{M}_{\text{ps+om}}$ denotes the model with correctly specified $p_z(X; \gamma)$ and $\mu_{zs}(X; \beta)$. We also use the union notation from the standard set theory to denote the correct specification of at least one model, for example, $\mathcal{M}_{\text{tp}} \cup \mathcal{M}_{\text{ps+om}}$ denotes the model with correctly specified $\pi(X; \alpha)$ or $\{p_z(X; \gamma), \mu_{zs}(X; \beta)\}$.

To obtain the estimators based on Theorem 1, we need to replace the components in the identification formulas with their estimated counterparts, and the expectations with the empirical averages. We use $\{\pi(X; \hat{\alpha}), p_z(X; \hat{\gamma}), \mu_{zs}(X; \hat{\beta})\}$ to denote the estimated version of $\{\pi(X; \alpha), p_z(X; \gamma), \mu_{zs}(X; \beta)\}$. For p_1 and p_0, we can simply use $\mathbb{P}_n\{p_1(X; \hat{\gamma})\}$ and $\mathbb{P}_n\{p_0(X; \hat{\gamma})\}$ as the estimators, which are consistent under $\mathcal{M}_{\text{ps}}$. Moreover, the doubly robust estimators (Bang & Robins, 2005),

$$\hat{p}_1 = \mathbb{P}_n\left[\frac{Z\{S - p_1(X; \hat{\gamma})\}}{\pi(X; \hat{\alpha})} + p_1(X; \hat{\gamma})\right], \quad \hat{p}_0 = \mathbb{P}_n\left[\frac{(1 - Z)\{S - p_0(X; \hat{\gamma})\}}{1 - \pi(X; \hat{\alpha})} + p_0(X; \hat{\gamma})\right],$$

improve them, which are consistent for p_1 and p_0 under $\mathcal{M}_{\text{tp}} \cup \mathcal{M}_{\text{ps}}$.

The identification formulas in Theorem 1(a) motivate the following weighting estimators based on the treatment probability and principal score.

Example 5 The treatment probability–principal score (tp-ps) estimators are

$$\hat{\tau}_{10,\mathrm{tp-ps}} = \mathbb{P}_n \left\{ \frac{e_{10}(X;\hat{\gamma})}{\hat{p}_1 - \hat{p}_0} \frac{S}{p_1(X;\hat{\gamma})} \frac{Z}{\pi(X;\hat{\alpha})} Y - \frac{e_{10}(X;\hat{\gamma})}{\hat{p}_1 - \hat{p}_0} \frac{1-S}{1-p_0(X;\hat{\gamma})} \frac{1-Z}{1-\pi(X;\hat{\alpha})} Y \right\},$$

$$\hat{\tau}_{00,\mathrm{tp-ps}} = \mathbb{P}_n \left\{ \frac{1-S}{1-\hat{p}_1} \frac{Z}{\pi(X;\hat{\alpha})} Y - \frac{e_{00}(X;\hat{\gamma})}{1-\hat{p}_1} \frac{1-S}{1-p_0(X;\hat{\gamma})} \frac{1-Z}{1-\pi(X;\hat{\alpha})} Y \right\},$$

$$\hat{\tau}_{11,\mathrm{tp-ps}} = \mathbb{P}_n \left\{ \frac{e_{11}(X;\hat{\gamma})}{\hat{p}_0} \frac{S}{p_1(X;\hat{\gamma})} \frac{Z}{\pi(X;\hat{\alpha})} Y - \frac{S}{\hat{p}_0} \frac{1-Z}{1-\pi(X;\hat{\alpha})} Y \right\}.$$

The weighting estimators in Example 5 involve the inverse of the treatment probability. Thus, they may be unstable if some estimated treatment probabilities are close to zero or one. A strategy to mitigate this issue is to stabilize the estimators by normalizing the weights (Hernán et al., 2001). For example, the stabilized weighting estimator of τ_{11} is

$$\hat{\tau}'_{11,\mathrm{tp-ps}} = \mathbb{P}_n \left\{ e_{11}(X;\hat{\gamma}) \frac{S}{p_1(X;\hat{\gamma})} \frac{Z}{\pi(X;\hat{\alpha})} Y \right\} \Big/ \mathbb{P}_n \left\{ e_{11}(X;\hat{\gamma}) \frac{S}{p_1(X;\hat{\gamma})} \frac{Z}{\pi(X;\hat{\alpha})} \right\}$$
$$- \mathbb{P}_n \left\{ \frac{s(1-Z)}{1-\pi(X;\hat{\alpha})} Y \right\} \Big/ \mathbb{P}_n \left\{ \frac{s(1-Z)}{1-\pi(X;\hat{\alpha})} \right\}.$$

The stabilized weighting estimators for τ_{10} and τ_{00} have similar forms. The estimators $\hat{\tau}_{u,\text{tp-ps}}$ are consistent under $\mathcal{M}_{\text{tp+ps}}$, that is, correct specifications of the treatment probability and principal score. However, if either model is incorrectly specified, they are inconsistent.

The identification formulas in Theorem 1(b) motivate the following estimators based on the treatment probability and the outcome mean.

Example 6 The treatment probability–outcome mean (tp-om) estimators are

$$\hat{\tau}_{10,\text{tp-om}} = \mathbb{P}_n \left[\frac{ZS/\pi(X;\hat{\alpha}) - (1-Z)S/\{1-\pi(X;\hat{\alpha})\}}{\hat{p}_1 - \hat{p}_0} \left\{ \mu_{11}(X;\hat{\beta}) - \mu_{00}(X;\hat{\beta}) \right\} \right],$$

$$\hat{\tau}_{00,\text{tp-om}} = \mathbb{P}_n \left[\frac{Z(1-S)/\pi(X;\hat{\alpha})}{1-\hat{p}_1} \left\{ \mu_{10}(X;\hat{\beta}) - \mu_{00}(X;\hat{\beta}) \right\} \right],$$

$$\hat{\tau}_{11,\text{tp-om}} = \mathbb{P}_n \left[\frac{(1-Z)S/\{1-\pi(X;\hat{\alpha})\}}{\hat{p}_0} \left\{ \mu_{11}(X;\hat{\beta}) - \mu_{01}(X;\hat{\beta}) \right\} \right].$$

Similar to the estimators in Example 5, we can also obtain the stabilized weighting version of the estimators in Example 6. For example, the stabilized version of $\hat{\tau}_{11,\text{tp-om}}$ is

$$\hat{\tau}'_{11,\text{tp-om}} = \mathbb{P}_n \left[\frac{(1-Z)S}{1-\pi(X;\hat{\alpha})} \left\{ \mu_{11}(X;\hat{\beta}) - \mu_{01}(X;\hat{\beta}) \right\} \right] \Big/ \mathbb{P}_n \left\{ \frac{(1-Z)S}{1-\pi(X;\hat{\alpha})} \right\}.$$

The estimators $\hat{\tau}_{u,\text{tp-om}}$ are consistent under $\mathcal{M}_{\text{tp+om}}$.

The identification formulas in Theorem 1(c) motivate the following estimators based on the principal score and outcome mean.

Example 7 The principal score–outcome mean (ps-om) estimators are

$$\hat{\tau}_{10,\mathrm{ps-om}} = \mathbb{P}_n \left[\frac{p_1(X;\hat{\gamma}) - p_0(X;\hat{\gamma})}{\hat{p}_1 - \hat{p}_0} \left\{ \mu_{11}(X;\hat{\beta}) - \mu_{00}(X;\hat{\beta}) \right\} \right],$$

$$\hat{\tau}_{00,\mathrm{ps-om}} = \mathbb{P}_n \left[\frac{1-p_1(X;\hat{\gamma})}{1-\hat{p}_1} \left\{ \mu_{10}(X;\hat{\beta}) - \mu_{00}(X;\hat{\beta}) \right\} \right],$$

$$\widehat{\tau}_{11,\text{ps-om}} = \mathbb{P}_n\left[\frac{p_0(X;\widehat{\gamma})}{\widehat{p}_0}\left\{\mu_{11}(X;\widehat{\beta}) - \mu_{01}(X;\widehat{\beta})\right\}\right].$$

The estimators $\widehat{\tau}_{u,\text{ps-om}}$ are consistent under $\mathcal{M}_{\text{ps+om}}$.

4 | FROM THE EIFS TO TRIPLY ROBUST ESTIMATORS

Theorem 1 presents three identification formulas, which motivate infinitely many estimators for each PCE. This calls for the construction of more principled estimators. In this section, we derive the EIF for each PCE to motivate a new estimator. The EIFs below are derived under the nonparametric model of the observed-data distribution, which is a standard strategy in the literature. In particular, the derivation ignores the restrictions implied by the monotonicity assumption (cf. Frölich, 2007; Hong & Nekipelov, 2010). For simplicity, we use the terminology 'EIF' throughout.

4.1 | EIFs and the resulting estimators

Because the PCEs have a ratio form $\tau_u = \mathbb{E}\{(Y_1 - Y_0)\mathbf{1}(U = u)\}/\mathbb{P}(U = u)$, we will first define a general quantity to represent the EIFs of the numerators and denominators, and then combine them to have the EIFs for the PCEs.

Define the following quantity for any function $f(Y, S, X)$:

$$\psi_{f(Y_z,S_z,X)} = \frac{\mathbf{1}(Z = z)[f(Y,S,X) - \mathbb{E}\{f(Y,S,X)|X,Z = z\}]}{\mathbb{P}(Z = z|X)} + \mathbb{E}\{f(Y,S,X)|X,Z = z\}. \tag{5}$$

Under Assumption 1, we can show that $\mathbb{E}\{\psi_{f(Y_z,S_z,X)}\} = \mathbb{E}\{f(Y_z,S_z,X)\}$. In fact, $\psi_{f(Y_z,S_z,X)} - \mathbb{E}\{f(Y_z,S_z,X)\}$ is the EIF for $\mathbb{E}\{f(Y_z,S_z,X)\}$; see Lemma S5 in the supplementary material. With $f(Y, S, X) = S$, Equation (5) reduces to

$$\psi_{S_z} = \frac{\mathbf{1}(Z = z)\{S - p_z(X)\}}{\mathbb{P}(Z = z|X)} + p_z(X),$$

and $\psi_{S_z} - \mathbb{E}(S_z)$ is the EIF for $\mathbb{E}(S_z)$. This reduces to a standard result in observational studies (Hahn, 1998), which is the foundation for constructing the doubly robust estimator for $\mathbb{E}(S_z)$ (Bang & Robins, 2005). With $f(Y, S, X) = YS$ and $z = 0$, Equation (5) reduces to

$$\psi_{Y_0S_0} = \frac{\mathbf{1}(Z = 0)\{YS - \mu_{01}(X)p_0(X)\}}{1 - \pi(X)} + \mu_{01}(X)p_0(X),$$

and $\psi_{Y_0S_0} - \mathbb{E}(Y_0S_0)$ is the EIF for $\mathbb{E}(Y_0S_0)$, which equals $\mathbb{E}(Y_0|U = 11)\mathbb{P}(U = 11)$ because $S_0 = 1$ is equivalent to $U = 11$ under monotonicity. Based on the ψ notation in Equation (5), the following theorem gives the EIFs for the PCEs.

Theorem 2 (EIFs). *Suppose τ_u's are identified in Theorem 1. The EIF for τ_{10} is $\phi_{10} = \{\phi_{1,10} - \phi_{0,10} - \tau_{10}(\psi_{S_1} - \psi_{S_0})\}/(p_1 - p_0)$, where*

$$\phi_{1,10} = \frac{e_{10}(X)}{p_1(X)}\psi_{Y_1S_1} - \mu_{11}(X)\left\{\psi_{S_0} - \frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\},$$

$$\phi_{0,10} = \frac{e_{10}(X)}{1-p_0(X)}\psi_{Y_0(1-S_0)} - \mu_{00}(X)\left\{\psi_{1-S_1} - \frac{1-p_1(X)}{1-p_0(X)}\psi_{1-S_0}\right\}.$$

The EIF for τ_{00} *is* $\phi_{00} = \left(\phi_{1,00} - \phi_{0,00} - \tau_{00}\psi_{1-S_1}\right)/(1-p_1)$, *where*

$$\phi_{1,00} = \psi_{Y_1(1-S_1)}, \quad \phi_{0,00} = \frac{e_{00}(X)}{1-p_0(X)}\psi_{Y_0(1-S_0)} + \mu_{00}(X)\left\{\psi_{1-S_1} - \frac{1-p_1(X)}{1-p_0(X)}\psi_{1-S_0}\right\}.$$

The EIF for τ_{11} is $\phi_{11} = \left(\phi_{1,11} - \phi_{0,11} - \tau_{11}\psi_{S_0}\right)/p_0$, *where*

$$\phi_{1,11} = \frac{e_{11}(X)}{p_1(X)}\psi_{Y_1S_1} + \mu_{11}(X)\left\{\psi_{S_0} - \frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\}, \quad \phi_{0,11} = \psi_{Y_0S_0}.$$

From Theorem 2, the semiparametric efficiency bounds for the PCEs are $\mathbb{E}(\phi_u^2)$ for $u = 10, 00, 11$ (Bickel et al., 1993). The EIFs have mean zero, so we can obtain another set of identification formulas by solving $\mathbb{E}(\phi_u) = 0$.

Corollary 1 *Under Assumptions* 1–3,

$$\tau_{10} = \frac{\mathbb{E}(\phi_{1,10} - \phi_{0,10})}{\mathbb{E}(\psi_{S_1} - \psi_{S_0})}, \quad \tau_{00} = \frac{\mathbb{E}(\phi_{1,00} - \phi_{0,00})}{\mathbb{E}(1-\psi_{S_1})}, \quad \tau_{11} = \frac{\mathbb{E}(\phi_{1,11} - \phi_{0,11})}{\mathbb{E}(\psi_{S_0})}. \tag{6}$$

As a sanity check of Equation (6), we can verify that the denominator of τ_u in Equation (6) equals $\mathbb{P}(U = u)$, and the numerator equals $\mathbb{E}\{(Y_1 - Y_0)\mathbf{1}(U = u)\}$, for $u = 10, 00, 11$. Based on Corollary 1, we can improve the estimators in Examples 5–7. Denote the estimator for $\psi_{f(Y_z,S_z,X)}$ by

$$\widehat{\psi}_{f(Y_z,S_z,X)} = \frac{\mathbf{1}(Z=z)[f(Y,S,X) - \widehat{\mathbb{E}}\{f(Y,S,X)|X,Z=z\}]}{\pi^z(X;\widehat{\alpha})\{1-\pi(X;\widehat{\alpha})\}^{1-z}} + \widehat{\mathbb{E}}\{f(Y,S,X)|X,Z=z\},$$

where $\widehat{\mathbb{E}}\{f(Y,S,X)|X,Z=z\}$ is the fitted conditional expectation of $f(Y,S,X)$ given X and $Z=z$. When $f(Y,S,X) = S$, we have $\widehat{\mathbb{E}}\{f(Y,S,X)|X,Z=z\} = p_z(X;\widehat{\gamma})$, which reduces to the estimated principal score and results in the estimator $\mathbb{P}_n(\widehat{\psi}_{S_z}) = \widehat{p}_z$. When $f(Y,S,X) = YS$, we have $\widehat{\mathbb{E}}\{f(Y,S,X)|X,Z=z\} = \mu_{z1}(X;\widehat{\beta})p_z(X;\widehat{\gamma})$, which relies on both the principal score and outcome mean.

Corollary 1 motivates the following estimators:

$$\widehat{\tau}_{10} = \frac{\mathbb{P}_n(\widehat{\phi}_{1,10} - \widehat{\phi}_{0,10})}{\mathbb{P}_n(\widehat{\psi}_{S_1} - \widehat{\psi}_{S_0})}, \quad \widehat{\tau}_{00} = \frac{\mathbb{P}_n(\widehat{\phi}_{1,00} - \widehat{\phi}_{0,00})}{\mathbb{P}_n(1-\widehat{\psi}_{S_1})}, \quad \widehat{\tau}_{11} = \frac{\mathbb{P}_n(\widehat{\phi}_{1,11} - \widehat{\phi}_{0,11})}{\mathbb{P}_n(\widehat{\psi}_{S_0})}, \tag{7}$$

where

$$\widehat{\phi}_{1,10} = \frac{e_{10}(X;\widehat{\gamma})}{p_1(X;\widehat{\gamma})}\widehat{\psi}_{Y_1S_1} - \mu_{11}(X;\widehat{\beta})\left\{\widehat{\psi}_{S_0} - \frac{p_0(X;\widehat{\gamma})}{p_1(X;\widehat{\gamma})}\widehat{\psi}_{S_1}\right\},$$

$$\widehat{\phi}_{0,10} = \frac{e_{10}(X;\widehat{\gamma})}{1-p_0(X;\widehat{\gamma})}\widehat{\psi}_{Y_0(1-S_0)} - \mu_{00}(X;\widehat{\beta})\left\{\widehat{\psi}_{1-S_1} - \frac{1-p_1(X;\widehat{\gamma})}{1-p_0(X;\widehat{\gamma})}\widehat{\psi}_{1-S_0}\right\},$$

$$\widehat{\phi}_{0,00} = \frac{e_{00}(X;\widehat{\gamma})}{1-p_0(X;\widehat{\gamma})}\widehat{\psi}_{Y_0(1-S_0)} + \mu_{00}(X;\widehat{\beta})\left\{\widehat{\psi}_{1-S_1} - \frac{1-p_1(X;\widehat{\gamma})}{1-p_0(X;\widehat{\gamma})}\widehat{\psi}_{1-S_0}\right\},$$

$$\hat{\phi}_{1,11} = \frac{e_{11}(X;\hat{\gamma})}{p_1(X;\hat{\gamma})}\hat{\psi}_{Y_1S_1} + \mu_{11}(X;\hat{\beta})\left\{\hat{\psi}_{S_0} - \frac{p_0(X;\hat{\gamma})}{p_1(X;\hat{\gamma})}\hat{\psi}_{S_1}\right\},$$
$$\hat{\phi}_{1,00} = \hat{\psi}_{Y_1(1-S_1)},$$
$$\hat{\phi}_{0,11} = \hat{\psi}_{Y_0S_0}.$$

These estimators for the PCEs are all in ratio forms, similar to the classic Wald estimator for the complier average causal effect under the monotonicity and ER (Angrist et al., 1996).

Motivating estimators based on EIFs is a standard approach in semiparametric statistics. This approach, however, involves advanced statistical theory. To add more intuition for the estimators above, we offer an alternative perspective in the supplementary material based on model-assisted estimation from the classic survey sampling theory. This extends the results on doubly robust and model-assisted estimation for the average causal effect in unconfounded observational studies (Kang & Schafer, 2007; Little & An, 2004; Lumley et al., 2011; Robins & Rotnitzky, 1998).

Interestingly, although the $\hat{\tau}_u$'s involve models for the treatment probability, principal score, and outcome, their consistency does not require the correct specification of all three models. We will characterize this *triple robustness* property in the next subsection.

4.2 | Triple robustness

The following theorem shows the triple robustness and local efficiency of the estimators constructed based on the EIFs.

Theorem 3 (Triple robustness and local efficiency). *Suppose that Assumptions 1–3 hold, $\delta < \{\pi(x;\alpha^*), \pi(x;\hat{\alpha})\} < 1-\delta$, and $\{p_1(x;\gamma^*), p_1(x;\hat{\gamma}), 1-p_0(x;\gamma^*), 1-p_0(x;\hat{\gamma})\} > \delta$ for some $\delta \in (0,1)$ and all x in the support of X. Each estimator $\hat{\tau}_u$ in Equation (7) is triply robust in the sense that it is consistent for τ_u under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$. Moreover, $\hat{\tau}_u$ has the influence function ϕ_u and therefore achieves the semiparametric efficiency bound under $\mathcal{M}_{\text{tp+ps+om}}$.*

The regularity condition in Theorem 3 is similar to the classic overlap condition (D'Amour et al., 2020; Rosenbaum & Rubin, 1983), which rules out small quantities in the denominators of the estimators. Theorem 3 states that $\hat{\tau}_u$ is consistent if any two of the three models are correctly specified, and locally efficient if all three models are correctly specified. For the variance calculation of these estimators, we use the nonparametric bootstrap.

To gain more intuition, we then give the sketch of the proof for the triple robustness of $\hat{\tau}_{10} = (\mathbb{P}_n\hat{\phi}_{1,10} - \mathbb{P}_n\hat{\phi}_{0,10})/(\mathbb{P}_n\hat{\psi}_{S_1} - \mathbb{P}_n\hat{\psi}_{S_0})$ and relegate additional technical details to the supplementary material. For simplicity in this paragraph, let $\mathcal{M}_{\text{triple}} = \mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$ denote the set containing at least two correct models. The denominator is consistent for $\mathbb{E}(S_1 - S_0) = \mathbb{P}(U=10)$ under $\mathcal{M}_{\text{tp}} \cup \mathcal{M}_{\text{ps}} \supseteq \mathcal{M}_{\text{triple}}$. For the terms in the numerator, we calculate their asymptotic biases in Section S7. In particular, $\mathbb{P}_n\hat{\phi}_{1,10} - \mathbb{E}\{Y_1\mathbf{1}(U=10)\}$ has the probability limit $B_1 + B_2 - B_3$, where

$$B_1 = \mathbb{E}\left[\frac{\{\mu_{11}(X)p_1(X) - \mu_{11}(X;\beta^*)p_1(X;\beta^*)\}\{\pi(X) - \pi(X;\alpha^*)\}}{\pi(X;\alpha^*)}\right],$$
$$B_2 = \mathbb{E}\left[\frac{\{\pi(X)p_0(X;\gamma^*)p_1(X) - \pi(X;\alpha^*)p_0(X)p_1(X;\gamma^*)\}\{\mu_{11}(X) - \mu_{11}(X;\beta^*)\}}{\pi(X;\alpha^*)p_1(X;\gamma^*)}\right],$$

$$B_3 = \mathbb{E}\left[\frac{\{\pi(X) - \pi(X;\alpha^*)\}\{p_0(X) - p_0(X;\gamma^*)\}\mu_{11}(X;\beta^*)}{1 - \pi(X;\alpha^*)}\right].$$

The bias B_1 equals 0 under $\mathcal{M}_{\text{ps+om}} \cup \mathcal{M}_{\text{tp}}$; the bias B_2 equals 0 under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{om}}$; the bias B_3 equals 0 under $\mathcal{M}_{\text{tp}} \cup \mathcal{M}_{\text{ps}}$. Each of these three sets contains $\mathcal{M}_{\text{triple}}$. As a result, $\mathbb{P}_n(\widehat{\phi}_{1,10})$ is consistent for $\mathbb{E}\{Y_1\mathbf{1}(U = 10)\}$ under $\mathcal{M}_{\text{triple}}$. Similarly, we can show $\mathbb{P}_n(\widehat{\phi}_{0,10})$ is consistent for $\mathbb{E}\{Y_0\mathbf{1}(U = 10)\}$ under $\mathcal{M}_{\text{triple}}$. So the triple robustness of $\widehat{\tau}_{10}$ holds.

The bias formulas above suggest that the proposed triply robust estimator would remain consistent and asymptotically Normal under some regularity conditions when using nonparametric or machine learning estimation for the nuisance functions $\pi(X)$, $p_z(X)$, and $\mu_{zs}(X)$, denoted by $\widehat{\pi}(X)$, $\widehat{p}_z(X)$, and $\widehat{\mu}_{zs}(X)$. This property would be similar to that of the doubly robust estimator for estimating the average causal effect in unconfounded observational studies (Chernozhukov et al., 2018). In other contexts involving intermediate variables, Zheng and van der Laan (2017) and Miles et al. (2020) have established similar results for multiply robust estimators. Theorem 4 formalizes the results for the proposed estimators using nonparametric or machine learning estimation.

Theorem 4 (Triple machine learning estimation). *Suppose that Assumptions 1–3 hold,*

(a) $\{\widehat{\pi}(x), \widehat{p}_z(x), \widehat{\mu}_{zs}(x)\} \to \{\pi(x), p_z(x), \mu_{zs}(x)\}$ *in probability for all x in the support of X,*
(b) $\{\widehat{\pi}(x), \widehat{p}_z(x), \widehat{\mu}_{zs}(x)\}$ and $\{\pi(x), p_z(x), \mu_{zs}(x)\}$ *are in a Donsker class,*
(c) $\delta < \{\pi(x), \widehat{\pi}(x)\} < 1 - \delta$, $\{p_1(x), \widehat{p}_1(x), 1 - p_0(x), 1 - \widehat{p}_0(x)\} > \delta$ *and* $\{|\widehat{\mu}_{zs}(x)|, |\mu_{zs}(x)|\} < C$ *for some* $\delta \in (0, 1)$, $C > 0$, *and all x in the support of X, and*
(d) $\|\widehat{g}(X) - g(X)\|_2 \times \|\widehat{h}(X) - h(X)\|_2 = o_{\mathbb{P}}(n^{-1/2})$, *for any* $g \neq h \in (\pi, p_z, \mu_{zs})$, *where* $\|\cdot\|_2$ *denotes the L_2-norm, that is,* $\|f(X)\|_2^2 = \int f(x)^2 \mathrm{d}F_X(x)$.

Then $\widehat{\tau}_u$ in Equation (5) is asymptotically Normal, has the influence function ϕ_u, and achieves the semiparametric efficiency bound.

Conditions (a)–(d) are analogous to those for double machine learning estimation of average causal effects (e.g. Bradic et al., 2019; Kennedy, 2016). The consistency in (a) and the rates of convergence in (d) are well studied for commonly used flexible models. Condition (b) restricts the complexity of the spaces of the nuisance functions and their estimators. The cross-fitting technique can be used to relax this condition (Chernozhukov et al., 2018). The conditions in (c) may not be necessary but enable bounding the error $|\widehat{\tau}_u - \mathbb{P}_n\phi_u|$ by the summation of the terms in the form of $\|\widehat{g}(X) - g(X)\|_2 \times \|\widehat{h}(X) - h(X)\|_2$ with $g \neq h \in (\pi, p_z, \mu_{zs})$. Thus, by Condition (d), the results in Theorem 4 follow.

Section S3 in the supplementary material extends the identification and estimation framework to two important scenarios under randomization, that is, $Z \perp\!\!\!\perp (S_1, S_0, Y_1, Y_0, X)$, and strong monotonicity, that is, $S_1 \geq S_0$ respectively. We also establish robustness properties of the corresponding estimators there.

5 | SIMULATION

We evaluate the finite-sample properties of various estimators at sample size $n = 500$. Generate covariate $X \in \mathbb{R}^5$ by $X_j \sim \mathrm{N}(0.25, 1)$ for $j = 1, \ldots, 4$, and $X_5 \sim$ Bernoulli (0.5). We use linear predictors, $C_j = X_j - 0.25$, or quadratic predictors, $\widetilde{C}_j = (X_j^2 - 1)/\sqrt{2}$, for $j = 1, \ldots, 4$. Generate the treatment by $Z|X \sim \text{Bernoulli}\{\pi(X)\}$, the intermediate variable by $S|(Z=z,X) \sim \text{Bernoulli}\{p_z(X)\}$, and the outcome by $Y|(Z=z,S=s,X) \sim \text{N}\{\mu_{zs}(X),1\}$. To assess the robustness of the estimators to model misspecification, we consider two different choices for each of $\pi(X)$, $p_z(X)$ and $\mu_{zs}(X)$, summarized in Table 2. We indicate the models by the name of the dependent variable and whether or not the predictors are linear. For example, 'tp:no' is the model with $\pi(X)=2\sum_{j=1}^{4}\tilde{C}_j/5$, and 'ps:yes' is the model with $p_z(X)=2\{(2z-1)-\sum_{j=1}^{4}C_j\}/5$.

We calculate the true value of τ_u based on the identification formulas in Theorem 1 and the true models. We then compare the following estimators for τ_u:

1. weighting estimators: $\hat{\tau}_{u,\text{tp-ps}}$ and $\hat{\tau}'_{u,\text{tp-ps}}$ given in Example 5;
2. regression estimators: $\hat{\tau}_{u,\text{tp-om}}$ given in Example 6 and $\hat{\tau}_{u,\text{ps}-\text{om}}$ given in Example 7;
3. triply robust estimators: $\hat{\tau}_u$ and $\hat{\tau}_{u,\text{ml}}$ with parametric models and with flexible generalized additive models for nuisance functions respectively.

We also consider the weighting estimator and the regression estimator in Ding and Lu (2017), which are proposed under randomized experiments. Under 'ps:yes' and 'ps:no', we estimate the principal score by logistic regressions with linear predictors X and (X_1,X_2) respectively; we estimate the outcome mean by linear regressions with the linear predictor X. Therefore, under generative models with the label 'yes', the fitting models are correctly specified, while under generative models with the label 'no', the fitting models are misspecified.

We compare the estimators in $2^3=8$ scenarios depending on whether the treatment probability, principal score or outcome model is correctly specified. Figure 1 presents the violin plots based on 1000 repeated sampling of the estimators. For all the three PCEs, the weighting estimators $\hat{\tau}_{u,\text{tp-ps}}$ and $\hat{\tau}'_{u,\text{tp-ps}}$ (indicated by 'w1' and 'w2' in the figures) are biased when the treatment probability or principal score model is misspecified. The bias with a misspecified treatment probability is larger than that with a misspecified principal score because the weights corresponding to the treatment probability are unbounded while the weights corresponding to the principal score are bounded within [0, 1]. The weighting estimator in Ding and Lu (2017) (indicated by 'w3' in the figures) performs similarly to $\hat{\tau}_{u,\text{tp-ps}}$ and $\hat{\tau}'_{u,\text{tp-ps}}$, because the treatment is randomized under 'tp:yes'. As our theory predicts, the regression estimator $\hat{\tau}_{u,\text{tp-om}}$ (indicated by 'r1' in the figures) is unbiased under $\mathcal{M}_{\text{tp+om}}$; the regression estimator $\hat{\tau}_{u,\text{ps}-\text{om}}$ (indicated by 'r2' in the figures) is unbiased under $\mathcal{M}_{\text{ps+om}}$. The regression estimator in Ding and Lu (2017) (indicated by 'r3' in the figures) performs similarly to $\hat{\tau}_{u,\text{tp-om}}$ in terms of bias. The triply robust estimator $\hat{\tau}_u$ (indicated by 'tr' in the figures) is unbiased under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$, verifying its triple robustness. With flexible models for the nuisance functions, $\hat{\tau}_{u,\text{ml}}$ is unbiased under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$ and is less biased than other estimators in most scenarios.

TABLE 2 Models for simulation with two specifications for each of logit$\{\pi(X)\}$, logit$\{p_z(X)\}$, and $\mu_{zs}(X)$, indicated by 'Yes' and 'No'

	logit$\{\pi(X)\}$	**logit$\{p_z(X)\}$**	**$\mu_{zs}(X)$**
Yes	0	$2\{(2z-1)-\sum_{j=1}^{4}C_j\}/5$	$\sum_{j=1}^{5}C_j(1+z+s)/4$
No	$2\sum_{j=1}^{4}\tilde{C}_j/5$	$2\{(2z-1)-\sum_{j=1}^{4}\tilde{C}_j\}/5$	$\sum_{j=1}^{5}\tilde{C}_j(1+z+s)/4$

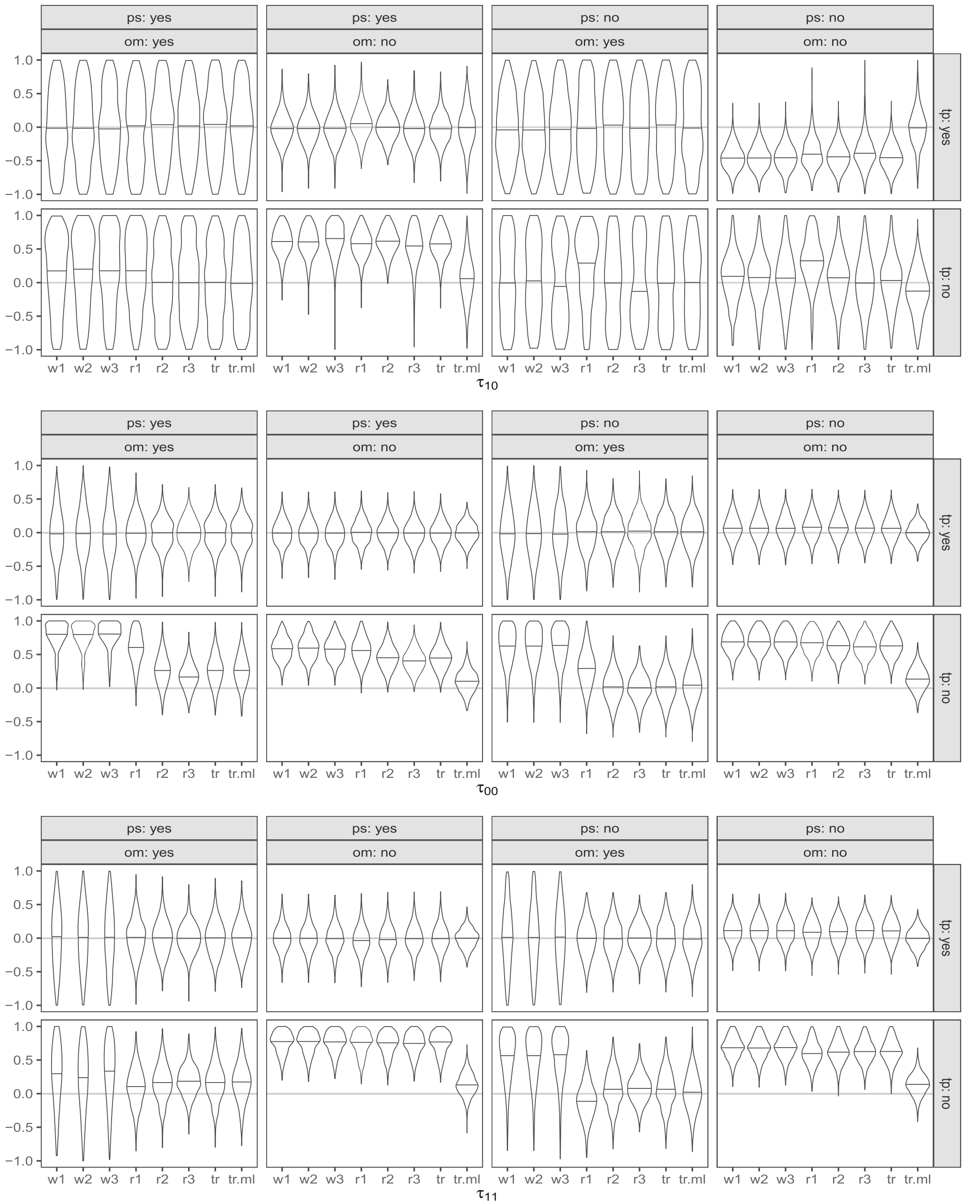

FIGURE 1 Violin plots of estimators in eight scenarios. Labels: 'w1' for $\hat{\tau}_{u,\text{tp-ps}}$, 'w2' for $\hat{\tau}'_{u,\text{tp-ps}}$, and 'w3' for the weighting estimator in Ding and Lu (2017); 'r1' for $\hat{\tau}_{u,\text{tp-om}}$; 'r2' for $\hat{\tau}_{u,\text{ps}-\text{om}}$; 'r3' for the regression estimator in Ding and Lu (2017); 'tr' for the triply robust estimator $\hat{\tau}_u$ and 'tr.ml' for the triply robust estimator $\hat{\tau}_{u,\text{ml}}$

6 | APPLICATIONS TO TWO OBSERVATIONAL STUDIES

6.1 | Return to schooling

The dataset from the U.S. National Longitudinal Survey of Young Men contains 3010 men with age between 14 and 24 in the year 1966. Card (1993) uses it to estimate the causal effect of education on earnings, utilizing the geographic variation in college proximity as an instrumental variable for education. Thus, the treatment Z is the indicator of growing up near a 4-year college; the intermediate variable S is the indicator of whether the individual receives education beyond high school; and the outcome Y is the log wage in the year 1976, ranging from 4.6 to 7.8. Monotonicity is plausible because living close to a college would make an individual more likely to receive higher education. To make principal ignorability plausible, we include the following covariates: race, age and squared age, a categorical variable indicating living with both parents, single mom, or both parents, and variables summarizing the living areas in the past. Unlike Card (1993), we do not invoke the ER that living near a college affected the earnings only through education. Rather, under principal ignorability, our analysis can assess the plausibility of the ER. Guo et al. (2014) and Yang et al. (2014) used similar strategies to test the ER.

We use a linear model for the outcome mean and logistic models for the treatment probability and principal score, and estimate the asymptotic variances by the nonparametric bootstrap. Table 3 presents the results for the estimated proportions of principal strata ($\hat{e}_u$) and PCEs using the weighting estimators ($\hat{\tau}_{u,\text{tp-ps}}$ and $\hat{\tau}'_{u,\text{tp-ps}}$), regression estimators ($\hat{\tau}_{u,\text{tp-om}}$ and $\hat{\tau}_{u,\text{ps}-\text{om}}$), and the triply robust estimator ($\hat{\tau}_u$). We omit $\hat{\tau}_{u,\text{tr.ml}}$ because it produces similar results as $\hat{\tau}_u$. All estimators are close, except for the unstabilized weighting estimator. This is due to the extreme fitted treatment probabilities. The estimators for τ_{00} and τ_{11} are not significant, suggesting no significant evidence of violating the ER. The estimated τ_{10} is positive and statistically significant, implying education has a positive effect on earnings. This finding corroborates with previous analyses.

6.2 | Causal effect of flooding on health

We re-analyse a dataset from Guo et al. (2018) with 774 households in Bangladesh to investigate the effect of flooding on children's diarrhoea. The treatment Z is the indicator of whether a household was severely affected by the flood; the intermediate variable S is the indicator of whether

TABLE 3 Analysis of the national longitudinal survey of young men (all significant effects are in bold)

	$u = 10$	$u = 00$	$u = 11$
$\hat{e}_u$	7% (3%, 10%)	48% (46%, 50%)	45% (42%, 49%)
$\hat{\tau}_{u,\text{tp-ps}}$	**−0.87** (**−1.69**, **−0.05**)	0.10 (−0.25, 0.44)	0.50 (−0.06, 1.05)
$\hat{\tau}'_{u,\text{tp-ps}}$	**0.15** (**0.00**, **0.30**)	0.01 (−0.04, 0.06)	0.02 (−0.04, 0.08)
$\hat{\tau}_{u,\text{tp-om}}$	0.09 (−0.03, 0.21)	0.02 (−0.02, 0.07)	0.01 (−0.05, 0.08)
$\hat{\tau}_{u,\text{ps}-\text{om}}$	**0.12** (**0.03**, **0.21**)	0.02 (−0.03, 0.07)	0.01(−0.05, 0.07)
$\hat{\tau}_u$	0.10 (−0.01, 0.23)	0.02 (−0.03, 0.07)	0.01 (−0.05, 0.07)

the per capita calorie consumption of the household was less than 2000 calories; and the outcome Y is the number of days a child had diarrhoea. Monotonicity is plausible because the calorie consumption would be negatively affected if the household was severely affected by the flood. To ensure principal ignorability, we include the following covariates: gender, age, the size of the household, mother's education, father's education, mother's age and father's age. As pointed out by Del Ninno (2001), the ER might be violated due to an alternative pathway through mother's health. We use our method to evaluate this assumption by estimating τ_{00} and τ_{11}.

Again we use a linear model for the outcome mean and logistic models for the treatment probability and principal score. Table 4 presents the results for the estimated proportions of principal strata and PCEs. The estimated τ_{00} and τ_{11} are both positive and the estimated τ_{00} is statistically significant, indicating that being affected by the flood tends to directly increase the number of days of diarrhoea. This also confirms the suspicion of the violation of the ER in Del Ninno (2001). Although the estimated τ_{10} is positive, it is imprecisely estimated and not statistically significant, due to the small proportion of stratum $U = 10$.

7 | DISCUSSION

PCEs characterize subgroup causal effects of important scientific meanings, providing insights into the underlying causal mechanism between the treatment and outcome. We develop an identification and estimation framework for PCEs under principal ignorability. The proposed estimators are analogous to those for the average causal effect in unconfounded observational studies. They are easy to implement which involve the model fitting of the treatment, intermediate variable and outcome conditional on baseline covariates. They are triply robust and locally efficient, naturally extending the classic doubly robust estimator for the average causal effect.

In mediation analysis, Tchetgen Tchetgen and Shpitser (2012) develop a general semiparametric framework for the direct and indirect effects. They focus on two scalar estimands, the natural direct and indirect effects, in the overall population. In contrast, we focus on the treatment effects within principal strata, resulting in more estimands in different subpopulations. Although the PCEs and the direct and indirect effects are related in certain scenarios (Forastiere et al., 2018; VanderWeele, 2008, 2011) there is no universal relationship between them, and the PCEs are applicable to a number of applications other than mediation analysis. Moreover, as discussed in Section 2, the identification assumptions in the two methods concern different aspects of the relationship between the potential outcome and the potential intermediate variable.

TABLE 4 Analysis of the flood data (all significant effects are in bold)

	$u = 10$	$u = 00$	$u = 11$
$\hat{e}_u$	9% (2%, 15%)	45% (41%, 50%)	46% (41%, 51%)
$\hat{\tau}_{u,\text{tp-ps}}$	0.74 (−3.60, 5.09)	**0.98** (**0.19**, **1.77**)	1.97 (−0.52, 2.47)
$\hat{\tau}'_{u,\text{tp-ps}}$	0.86 (−3.35, 5.07)	**0.92** (**0.13**, **1.71**)	1.11 (−0.36, 2.58)
$\hat{\tau}_{u,\text{tp-om}}$	1.74 (−3.55, 7.03)	**0.93** (**0.10**, **1.77**)	1.01 (−0.45, 2.47)
$\hat{\tau}_{u,\text{ps−om}}$	1.50 (−3.01, 6.01)	**0.92** (**0.09**, **1.75**)	1.10 (−0.33, 2.53)
$\hat{\tau}_u$	1.51 (−3.68, 6.71)	**0.88** (**0.05**, **1.71**)	1.10 (−0.33, 2.53)

We can generalize the theory to other causal estimands within principal strata by invoking a stronger version of principal ignorability: $Y_1 \perp\!\!\!\perp U|(Z=1,S=1,X)$ and $Y_0 \perp\!\!\!\perp U|(Z=0,S=0,X)$. Under this assumption, we can identify the effects on a transformation of the outcome: $\mathbb{E}\{h(Y_1,X)|U=u\} - \mathbb{E}\{h(Y_0,X)|U=u\}$ for any function $g(\cdot)$. This further ensures the identification of distributional or quantile causal effects within principal strata. Similar to the main paper, we can also derive EIFs and propose robust estimators for these causal estimands.

More generally, we can extend the results to deal with a continuous S, where the number of principal strata is infinity. The principal score becomes the conditional density of $U=(S_1,S_0)$ given covariates, which is not identifiable even under monotonicity. Therefore, the extension requires more sophisticated identification and estimation strategies. We leave the technical investigation to future research.

ACKNOWLEDGEMENTS

We thank the Associate Editor, three reviewers, Anqi Zhao and Sizhu Lu for helpful comments. Peng Ding was partially funded by the U.S. National Science Foundation (grant # 1945136).

DATA AVAILABILITY STATEMENT

The data used in our paper are publicly available.

ORCID

Zhichao Jiang https://orcid.org/0000-0002-8571-0217
Shu Yang https://orcid.org/0000-0001-7703-707X
Peng Ding https://orcid.org/0000-0002-2704-2353

REFERENCES

Abadie, A. (2003) Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113, 231–263.

Angrist, J.D. & Imbens, G.W. (1994) Identification and estimation of local average treatment effects. *Econometrica*, 62, 467–475.

Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996) Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 91, 444–455.

Bang, H. & Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

Bickel, P.J., Klaassen, C., Ritov, Y. & Wellner, J. (1993) *Efficient and adaptive inference in semiparametric models*. Baltimore, MD: Johns Hopkins University Press.

Bradic, J., Wager, S. & Zhu, Y. (2019) Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*.

Card, D. (1993) Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research.

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

D'Amour, A., Ding, P., Feller, A., Lei, L. & Sekhon, J. (2020) Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221, 644–654.

Del Ninno, C. (2001) *The 1998 floods in Bangladesh: disaster impacts, household coping strategies, and response*, Volume 122. Washington, D.C.: International Food Policy Research Institute.

Ding, P. & Lu, J. (2017) Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 757–777.

Ding, P., Geng, Z., Yan, W. & Zhou, X.-H. (2011) Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106, 1578–1591.

Egleston, B.L., Scharfstein, D.O. & MacKenzie, E. (2009) On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*, 65, 497–504.

Elliott, M.R., Raghunathan, T.E. & Li, Y. (2010) Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, 11, 353–372.

Feller, A., Mealli, F. & Miratrix, L. (2017) Principal score methods: assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics*, 42, 726–758.

Feller, A., Greif, E., Ho, N., Miratrix, L. & Pillai, N. (2022) Weak separation in mixture models and implications for principal stratification. In: *International Conference on Artificial Intelligence and Statistics*, in press.

Follmann, D.A. (2000) On the effect of treatment among would-be treatment compliers: an analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*, 95, 1101–1109.

Follmann, D. (2006) Augmented designs to assess immune response in vaccine trials. *Biometrics*, 62, 1161–1169.

Forastiere, L., Mattei, A. & Ding, P. (2018) Principal ignorability in mediation analysis: through and beyond sequential ignorability. *Biometrika*, 105, 979–986.

Frangakis, C.E. & Rubin, D.B. (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86, 365–379.

Frangakis, C.E. & Rubin, D.B. (2002) Principal stratification in causal inference. *Biometrics*, 58, 21–29.

Frölich, M. (2007) Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139, 35–75.

Frumento, P., Mealli, F., Pacini, B. & Rubin, D.B. (2012) Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107, 450–466.

Gallop, R., Small, D.S., Lin, J.Y., Elliott, M.R., Joffe, M.& Ten Have, T.R. (2009) Mediation analysis with principal stratification. *Statistics in Medicine*, 28, 1108–1130.

Gilbert, P.B. & Hudgens, M.G. (2008) Evaluating candidate principal surrogate endpoints. *Biometrics*, 64, 1146–1154.

Guo, Z., Cheng, J., Lorch, S.A. & Small, D.S. (2014) Using an instrumental variable to test for unmeasured confounding. *Statistics in Medicine*, 33, 3528–3546.

Guo, Z., Small, D.S., Gansky, S.A. & Cheng, J. (2018) Mediation analysis for count and zero inflated count data without sequential ignorability and its application in dental studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67, 371–394.

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

Hayden, D., Pauler, D.K. & Schoenfeld, D. (2005) An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*, 61, 305–310.

Hernán, M.A., Brumback, B. & Robins, J.M. (2001) Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96, 440–448.

Hill, J., Waldfogel, J. & Brooks-Gunn, J. (2002) Differential effects of high-quality child care. *Journal of Policy Analysis and Management*, 21, 601–627.

Hirano, K., Imbens, G.W., Rubin, D.B. & Zhou, X.-H. (2000) Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1, 69–88.

Hong, H. & Nekipelov, D. (2010) Semiparametric efficiency in nonlinear late models. *Quantitative Economics*, 1(2), 279–304.

Huang, Y. & Gilbert, P.B. (2011) Comparing biomarkers as principal surrogate endpoints. *Biometrics*, 67, 1442–1451.

Imai, K. (2008) Sharp bounds on the causal effects in randomized experiments with 'truncation-bydeath'. *Statistics and Probability Letters*, 78, 144–149.

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

Jiang, Z. & Ding, P. (2021) Identification of causal effects within principal strata using auxiliary variables. *Statistical Science*, 36, 493–508.

Jiang, Z., Ding, P. & Geng, Z. (2016) Principal causal effect identification and surrogate end point evaluation by multiple trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 829–848.

Jo, B. & Stuart, E.A. (2009) On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, 28, 2857–2875.

Jo, B., Stuart, E.A., MacKinnon, D.P. & Vinokur, A.D. (2011) The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, 46, 425–452.

Kang, J.D.Y. & Schafer, J.L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 523–539.

Kennedy, E.H. (2016) Semiparametric theory and empirical processes in causal inference. In: *Statistical causal inferences and their applications in public health research*, Springer. pp. 141–167.

Li, Y., Taylor, J.M.G. & Elliott, M.R. (2010) A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics*, 66, 523–531.

Little, R. & An, H. (2004) Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14, 949–968.

Lumley, T., Shaw, P.A. & Dai, J.Y. (2011) Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79, 200–220.

Mattei, A. & Mealli, F. (2011) Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 729–752.

Mattei, A., Li, F. & Mealli, F. (2013) Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *Annals of Applied Statistics*, 7, 2336–2360.

Mattei, A., Mealli, F. & Pacini, B. (2014) Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics*, 70, 278–288.

Mealli, F. & Pacini, B. (2013) Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *Journal of the American Statistical Association*, 108, 1120–1131.

Miles, C.H., Shpitser, I., Kanki, P., Meloni, S. & Tchetgen Tchetgen, E.J. (2020) On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding. *Biometrika*, 107, 159–172.

Ogburn, E.L., Rotnitzky, A. & Robins, J.M. (2015) Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 373–396.

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

Robins, J.M. & Rotnitzky, A. (1998) Discussion of 'Robust models in probability sampling' by Firth and Bennett. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 51–52.

Robins, J.M., Richardson, T.S. & Shpitser, I. (2020) An interventionist approach to mediation analysis. *arXiv preprint arXiv:2008.06019*.

Rosenbaum, P.R. (1984) The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147, 656–666.

Rosenbaum, P.R. & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rubin, D.B. (2004) Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31, 161–170.

Rubin, D.B. (2006) Causal inference through potential outcomes and principal stratification: application to studies with 'censoring' due to death. *Statistical Science*, 21, 299–309.

Schwartz, S.L., Li, F. & Mealli, F. (2011) A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*, 106, 1331–1344.

Shi, X., Miao, W., Nelson, J.C. & Tchetgen Tchetgen, E.J. (2020) Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 8, 521–540.

Sommer, A. & Zeger, S.L. (1991) On estimating efficacy from clinical trials. *Statistics in Medicine*, 10, 45–52.

Stuart, E.A. & Jo, B. (2015) Assessing the sensitivity of methods for estimating principal causal effects. *Statistical Methods in Medical Research*, 24, 657–674.

Tan, Z. (2006) Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101, 1607–1618.

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

VanderWeele, T.J. (2008) Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*, 78, 2957–2962.

VanderWeele, T.J. (2011) Principal stratification—uses and limitations. *The International Journal of Biostatistics*, 7, 1–14.

Wang, L. & Tchetgen Tchetgen, E. (2018) Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 531–550.

Wang, L., Richardson, T.S. & Zhou, X.-H. (2017) Causal analysis of ordinal treatments and binary outcomes under truncation by death. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 719–735.

Yang, F. & Small, D.S. (2016) Using post-outcome measurement information in censoring-by-death problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 299–318.

Yang, F., Zubizarreta, J.R., Small, D.S., Lorch, S. & Rosenbaum, P.R. (2014) Dissonant conclusions when testing the validity of an instrumental variable. *American Statistician*, 68, 253–263.

Zhang, J.L., Rubin, D.B. & Mealli, F. (2008) Evaluating the effects of job training programs on wages through principal stratification. *Advances in Econometrics*, 21, 117–145.

Zhang, J.L., Rubin, D.B. & Mealli, F. (2009) Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104, 166–176.

Zheng, W. & van der Laan, M. (2017) Longitudinal mediation analysis with time-varying mediators and exposures, with application to survival outcomes. *Journal of Causal Inference*, 5, 20160006.

Zigler, C.M. & Belin, T.R. (2012) A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics*, 68, 922–932.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Jiang, Z., Yang, S. & Ding, P. (2022) Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4), 1423–1445. Available from: https://doi.org/10.1111/rssb.12538

Supplementary Material

Section S1 establishes the balancing properties of the treatment probability and principal score, and discusses an alternative estimation strategy for them based on these balancing properties.

Section S2 provides an alternative motivation for the triply robust estimators based on model-assisted estimation in survey sampling.

Section S3 extends the identification and estimation framework for the PCEs to two scenarios under randomization and strong monotonicity, respectively.

Section S4 proposes sensitivity analyses for principal ignorability and monotonicity, respectively.

Section S5 proves the identification results, including Theorem 1 in the main text, Theorem S1 Section S1, and Theorems S5 and S8 in Section S4.

Section S6 proves the EIFs, including Theorem 2 in the main text, Theorems S2, S3, and S4 in Section S3, and Theorems S6 and S9 in Section S4.

Section S7 proves the multiple robustness and local efficiency, including Theorems 3 and 4 in the main text and Theorems S7 and S10 in Section S3.

S1 The role of covariate balancing

S1.1 Balancing properties

In observational studies, Rosenbaum and Rubin (1983) prove the balancing property of the treatment probability: the covariate distributions are the same in the treatment and control groups conditional on it or weighted by its inverse. Now we generalize this result to the balancing properties of both the treatment probability and principal score.

Theorem S1 (Balancing properties) *For any function $h(X)$ that has finite moments $\mathbb{E}\{h(X) \mid U=u\} < \infty$ for $u = 10, 00, 11$, the following three sets of balancing conditions hold.*

(a) Corresponding to stratum $U = 10$, we have

$$
\begin{aligned}
\mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}h(X)\right\} &= \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}h(X)\right\} \\
&= \mathbb{E}\left[\frac{SZ/\pi(X)-S(1-Z)/\{1-\pi(X)\}}{p_1-p_0}h(X)\right] \\
&= \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}h(X)\right\}.
\end{aligned}
$$

(b) Corresponding to stratum $U = 00$, we have

$$
\mathbb{E}\left\{\frac{(1-S)Z}{(1-p_1)\pi(X)}h(X)\right\} = \mathbb{E}\left\{\frac{1-p_1(X)}{1-p_1}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}h(X)\right\}
$$

$$
\begin{aligned}
&= \mathbb{E}\left\{\frac{1-SZ/\pi(X)}{1-p_1}h(X)\right\} \\
&= \mathbb{E}\left\{\frac{1-p_1(X)}{1-p_1}h(X)\right\}.
\end{aligned}
$$

(c) Corresponding to stratum $U = 11$, we have

$$
\mathbb{E}\left\{\frac{p_0(X)}{p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}h(X)\right\} = \mathbb{E}\left[\frac{S(1-Z)}{p_0\{1-\pi(X)\}}h(X)\right] = \mathbb{E}\left\{\frac{p_0(X)}{p_0}h(X)\right\}.
$$

Under Assumptions 1 and 2, the three formulas in (a)–(c) *equal* $\mathbb{E}\{h(X) \mid U = u\}$ *for* $u = 10, 00, 11$, *respectively.*

Theorem S1 generalizes the two-way balancing properties in Ding and Lu (2017) to the four-way balancing properties. The terms involving p_1 and p_0 cancel in the above formulas. Nevertheless, we keep them to aid interpretation.

For each principal stratum, the balancing conditions consist of the expectations of $h(X)$ in pseudo-populations defined by four different weights. Because all except one of the weights involve the principal score, we would still have all these balancing conditions even in randomized experiments. Therefore, Theorem S1 provides more balancing conditions than those in Ding and Lu (2017). Although these weights take different forms, the conditional expectations of them given X are the same. For example, based on the discussion under Theorem 1, the conditional expectations of the weights in Theorem S1(a) all equal $e_{10}(X)/e_{10}$. A similar discussion applies to the weights in Theorem S1(b) and (c).

The identities in Theorem S1 follow from the multiple expressions of $\mathbb{E}\{h(X) \mid U = u\}$. Because any functions of covariates are unaffected by the treatment within principal strata, the PCEs on $h(X)$ are all zero. Then, the first identity in each of the three formulas can be viewed as special cases of Theorem 1(a) with Y replaced by $h(X)$. For example, by replacing Y with $h(X)$ in Theorem 1(a), the two terms on the right-hand side both equal $\mathbb{E}\{h(X) \mid U = u\}$. This yields the first identity in Theorem S1(a). On the other hand, $\mathbb{E}\{h(X) \mid U = u\}$ is the expectation of $h(X)$ over the distribution of X within stratum $U = u$. Therefore, we can obtain the second and third identities in each of the three formulas by replacing the outcome mean difference with $h(X)$ in Theorem 1(b) and (c). Theorem S1(c) has one fewer identity than (a) and (b), because the identity obtained from Theorem 1(a) coincides with that obtained from Theorem 1(b). Although the heuristics come from Theorem 1, the formulas in Theorem S1 hold as long as the treatment probability and principal score are correctly specified; they do not require Assumptions 1–3. If Assumptions 1 and 2 hold, then we can connect them with the expectations of $h(X)$ within principal strata.

S1.2 Applications of balancing properties

First, the balancing properties in Theorem S1 are the theoretical foundation for model checking of the treatment probability and principal score without touching the outcome data. Theorem S1 holds with the true treatment probability and principal score. Therefore, any empirical violations of the balancing properties provide the basis for refuting the models of the treatment probability and principal score. Checking for these balance conditions does not involve modeling the outcome mean. With such a balance checking procedure, it is reasonable to favor the estimators that are consistent under $\mathcal{M}_{\text{tp+ps}}$ (e.g., $\widehat{\tau}_{u,\text{tp-ps}}$ and $\widehat{\tau}_u$) over those that are consistent only under $\mathcal{M}_{\text{om}}$ (e.g., $\widehat{\tau}_{u,\text{tp-om}}$ and $\widehat{\tau}_{u,\text{ps-om}}$).

Second, by choosing different $h(X)$ in Theorem S1, the balancing properties imply infinitely many moment conditions. This allows for an alternative approach to constructing more efficient estimators. In the context of missing data, Graham (2011) proposes a strategy to utilize all the moment conditions from balancing properties, which turns out to be equivalent to the approach based on the EIF. Similarly, for each PCE, we can leverage the moment conditions implied by Theorem S1 to improve the estimators in Examples 5–7. The resulting estimator will be the same as the proposed triply robust estimator.

Lastly, the balancing properties provide an alternative estimation strategy for $\pi(X;\alpha)$ and $p_z(X;\gamma)$ by constructing estimating equations based on the identities in Theorem S1. For example, the identities in Theorem S1(c) imply estimating equations

$$\mathbb{E}\left\{p_0(X;\gamma)\frac{S}{p_1(X;\gamma)}\frac{Z}{\pi(X;\alpha)}h(X)\right\} = \mathbb{E}\left\{\frac{\mathrm{s}(1-Z)}{1-\pi(X;\alpha)}h(X)\right\} = \mathbb{E}\left\{p_0(X;\gamma)h(X)\right\}$$

for any $h(X)$. The identities in Theorem S1(a) and (b) imply other estimating equations. We can choose a set of $h(X)$, and then solve the corresponding estimating equations to obtain estimators of α and γ. This strategy is similar to the covariate balancing propensity score (Imai and Ratkovic, 2014) for estimating the average causal effect in observational studies. We can also construct balancing weights for estimating the PCEs based on Theorem S1, similar to the strategies for estimating the average causal effect (Zubizarreta, 2015; Tan, 2020). We do not pursue this direction in the current paper and leave it to future work.

S2 Connection with model-assisted estimation

To gain more intuition about the EIFs and the identification formulas, we provide an interpretation of (6) based on the model-assisted estimation for the PCEs. In observational studies for the average causal effects, augmented inverse probability weighting and some model-assisted estimation strategies are equivalent, both leveraging outcome modeling to improve efficiency and robustness (Robins and Rotnitzky, 1998; Little and An, 2004; Kang and Schafer, 2007; Lumley et al., 2011). However, as far as we know, analogous connections are missing between multiply robust estimation and model-assisted estimation. As an extension, we make a modest contribution to show that this equivalence also holds between (6) and certain forms of the model-assisted estimation for the PCEs.

We focus on τ_{10}. It has the following ratio form:

$$\tau_{10} = \frac{\mathbb{E}\{Y_1 \mathbf{1}(U=10)\} - \mathbb{E}\{Y_0 \mathbf{1}(U=10)\}}{\mathbb{E}(S_1 - S_0)}. \tag{S1}$$

We first discuss the model-assisted estimation for the denominator $\mathbb{E}(S_1 - S_0)$. The potential outcome S_z decomposes into two terms based on its model given X, i.e., the principal score $p_z(X) = \mathbb{E}(S_z \mid X)$:

$$\mathbb{E}(S_z) = \mathbb{E}\{S_z - p_z(X)\} + \mathbb{E}\{p_z(X)\}, \quad (z = 0, 1). \tag{S2}$$

Applying the inverse probability weighting formula to $\mathbb{E}\{S_z - p_z(X)\}$ yields

$$\mathbb{E}(S_1) = \mathbb{E}\left[\frac{Z\{S - p_1(X)\}}{\pi(X)} + p_1(X)\right], \quad \mathbb{E}(S_0) = \mathbb{E}\left[\frac{(1-Z)\{S - p_0(X)\}}{1 - \pi(X)} + p_0(X)\right].$$

This leads to $\mathbb{E}(S_1 - S_0) = \mathbb{E}(\psi_{S_1} - \psi_{S_0})$, the model-assisted estimation formula for $\mathbb{E}(S_1 - S_0)$, which involves the treatment probability and principal score. Subtracting $p_z(X)$ from S_z reduces the variation in S_z, and thus the model-assisted estimator improves the efficiency over the simple weighting estimator. This recovers the classic doubly robust estimator (Bang and Robins, 2005). Similarly, $\mathbb{E}\{\psi_{f(Y_z, S_z, X)}\}$ is the model-assisted estimation formula for $\mathbb{E}\{f(Y_z, S_z, X)\}$ with the model of $f(Y_z, S_z, X)$.

We then discuss the model-assisted estimation for the numerator of (S1). Decompose it based on the outcome mean:

$$\mathbb{E}[\{Y_1 - \mu_{11}(X)\}\mathbf{1}(U=10)] + \mathbb{E}\{\mu_{11}(X)\mathbf{1}(U=10)\}. \tag{S3}$$

Applying the inverse probability weighting formula in Theorem 1(a) to the first term of (S3) yields

$$\mathbb{E}[\{Y_1 - \mu_{11}(X)\}\mathbf{1}(U=10)] \quad = \quad \mathbb{E}\left[e_{10}(X)\frac{S}{p_1(X)}\frac{Z}{\pi(X)}\{Y - \mu_{11}(X)\}\right];$$

applying the model-assisted estimation formula to the second term of (S3) yields

$$\mathbb{E}\{\mu_{11}(X)\mathbf{1}(U=10)\} = \mathbb{E}\{\mu_{11}(X)(S_1 - S_0)\} = \mathbb{E}\left\{\mu_{11}(X)(\psi_{S_1} - \psi_{S_0})\right\}.$$

Importantly, the quantities inside the above two expectations sum to $\phi_{1,10}$:

$$e_{10}(X)\frac{S}{p_1(X)}\frac{Z}{\pi(X)}\{Y-\mu_{11}(X)\}+\mu_{11}(X)(\psi_{S_1}-\psi_{S_0})=\phi_{1,10}, \tag{S4}$$

which proves that $\mathbb{E}(\phi_{1,10})$ is the model-assisted estimation formula for $\mathbb{E}\{Y_1\mathbf{1}(U=10)\}$. Similarly, $\mathbb{E}(\phi_{0,10})$ is the model-assisted estimation formula for $\mathbb{E}\{Y_0\mathbf{1}(U=10)\}$. As a result, the model-assisted estimation formula for τ_{10} coincides with the one derived in (6).

The above perspective explains the efficiency gain of the estimator $\hat{\tau}_u$ ($u=10,00,11$) over the simple weighting estimators. More surprisingly, they also improve robustness over all estimators in Examples 5–7, as shown in Section 4.2 in the main paper.

S3 Extensions

We extend the general identification and estimation framework for the PCEs in Section 4 to two important scenarios with different assumptions.

S3.1 Randomized experiments

In randomized experiments, a stronger version of Assumption 1 holds.

Assumption S1 (Randomization) $Z \perp\!\!\!\perp (S_1, S_0, Y_1, Y_0, X)$.

Under Assumption S1, the identification formulas in Theorem 1 hold by replacing $\pi(X)$ with π. Theorem 1(a) reduces to the identification formulas in Ding and Lu (2017). Moreover, Theorem 1(b) and (c) provide two additional identification formulas for each of the PCEs. Under randomization, the EIFs have simpler forms, as shown in the following theorem.

Theorem S2 (EIFs under randomization) *Suppose τ_u's are identified by the formulas in Theorem 1 with $\pi(X)$ replacing by π. Under Assumption S1, the EIFs for the PCEs are the same as those in Theorem 2 with $\pi(X)$ replaced by π.*

Theorem S2 demonstrates that knowing the treatment probability does not change the EIFs for the PCEs, which is similar to Hahn (1998)'s result that knowing the treatment probability does not change the EIF for the average causal effect in uncounfounded observational studies. We recommend adopting our estimators $\widehat{\tau}_u$ even in randomized experiments. The following theorem shows the double robustness property of these estimators, which is a direct application of Theorem 3.

Theorem S3 (Double robustness and local efficiency under randomization) *Suppose Assumptions 2, 3, and S1 hold, $\delta < \{\pi, \pi(x;\hat{\alpha})\} < 1-\delta$, and $\{p_1(x;\gamma^*), p_1(x;\widehat{\gamma}), 1-p_0(x;\gamma^*), 1-p_0(x;\widehat{\gamma})\} > \delta$*

for some $\delta \in (0,1)$ and all x in the support of X. The estimator $\widehat{\tau}_u$ $(u = 10, 00, 11)$ is doubly robust in the sense that it is consistent for τ_u under $\mathcal{M}_{\text{ps}} \cup \mathcal{M}_{\text{om}}$. Moreover, $\widehat{\tau}_u$ achieves the semiparametric efficiency bound under $\mathcal{M}_{\text{ps+om}}$.

In randomized experiments, the treatment probability is always correctly specified by including the null model and thus the triple robustness simplifies to the double robustness with respect to the principal score and outcome mean. Because the treatment probability does not depend on X under randomization, we can simplify $\widehat{\tau}_u$ by replacing the estimated propensity score with the estimated treated proportion. However, if $\mathcal{M}_{\text{ps+om}}$ does not hold, then using the estimated propensity score might improve the efficiency of the estimator. This has been pointed out by Shen et al. (2014) for the inverse probability weighting estimator of the average causal effect under randomization.

Ding and Lu (2017) propose model-assisted estimators for the PCEs. However, their estimators are not based on the EIFs and use only the principal score model, without using the outcome model. As a result, their estimators are neither doubly robust nor semiparametrically efficient. Even in randomized experiments, their estimators are suboptimal, and we propose improved estimators.

S3.2 Strong monotonicity

When strong monotonicity holds, we have stronger results below.

Assumption S2 (Strong monotonicity) *$S_0 = 0$.*

Assumption S2 implies $S_1 \geq S_0$ and is thus stronger than Assumption 2. Assumption S2 eliminates principal strata $U = 11$ and $U = 01$, and restricts $p_0(X) = p_0 = 0$. Therefore, there are only two principal strata $U = 10$ and $U = 00$. The identification formulas for τ_{10} and τ_{00} can be obtained by applying these restrictions in Theorem 1. The following theorem gives their EIFs and shows the triple robustness property of the corresponding estimators.

Theorem S4 (EIFs and triple robustness under strong monotonicity) *Suppose τ_{10} and τ_{00} are identified by the formulas in Theorem 1 with $p_0(X) = p_0 = 0$. Under Assumption S2, the EIFs for τ_{10} and τ_{00} are the same as those given in Theorem 2 with $p_0(X)$, p_0 and ψ_{S_0} set to 0. Suppose further that $\delta < \{\pi(x;\alpha^*), \pi(x;\hat{\alpha})\} < 1 - \delta$, and $\{p_1(x;\gamma^*), p_1(x;\widehat{\gamma})\} > \delta$ for some $\delta \in (0,1)$ and all x in the support of X. The estimator $\widehat{\tau}_u$ $(u = 10, 00)$ is triply robust in the sense that it is consistent for τ_u under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$. Moreover, $\widehat{\tau}_u$ achieves the semiparametric efficiency bound under $\mathcal{M}_{\text{tp+ps+om}}$.*

We use the same estimators $\widehat{\tau}_{10}$ and $\widehat{\tau}_{00}$ for τ_{10} and τ_{00}, because empirically the data would inform that $p_0(X;\widehat{\gamma}) = \widehat{p}_0 = 0$ and $\widehat{\psi}_{S_0} = 0$. The triple robustness of $\widehat{\tau}_{10}$ and $\widehat{\tau}_{00}$ is the same as that given in Theorem 3, except that the models of $p_0(X)$ and $\mu_{01}(X)$ are no longer needed.

S4 Sensitivity analysis

Assumptions 2 and 3 are crucial for the nonparametric identification of the PCEs. However, these two assumptions cannot be easily justified by prior knowledge and may be violated in some applications. Therefore, we extend the sensitivity analysis in Ding and Lu (2017) from randomized experiments to observational studies and derive the semiparametric efficiency theory for the sensitivity analysis.

S4.1 Sensitivity analysis for principal ignorability: method

Assumptions 3 is violated if there are latent confounders between the principal strata and outcome. For the sensitivity analysis, we assume the following tilting model:

$$\frac{E(Y_1 \mid U = 10, X)}{E(Y_1 \mid U = 11, X)} = \epsilon_1(X), \quad \frac{E(Y_0 \mid U = 10, X)}{E(Y_0 \mid U = 00, X)} = \epsilon_0(X). \tag{S5}$$

Under (S5), we can treat $\epsilon_1(X)$ and $\epsilon_0(X)$ as sensitivity parameters. If $\epsilon_1(X) = \epsilon_0(X) = 1$, then Assumption 3 holds. When the outcome is non-negative and $\epsilon_1(X)$ and $\epsilon_0(X)$ are constant not depending on X, (S5) effectively assumes log-linear models on the outcome on the latent U and X, similar to Scharfstein et al. (1999). Model (S5) also extends Ding and Lu (2017) by allowing the sensitivity parameters to depend on X, e.g., $\epsilon_z(X) = \exp(-X^{\mathrm{T}}\eta_z)$ for $z = 0, 1$. The following theorem establishes the nonparametric identification of the PCEs when the sensitivity parameters are known.

Theorem S5 *Under Assumptions 1, 2, and* (S5) *with known* $\epsilon_1(X)$ *and* $\epsilon_0(X)$, $e_u > 0$ *for* $u = 10, 00, 11$, *and* $0 < \pi(x) < 1$ *for all* x *in the support of* X. *Define*

$$\begin{aligned}
\omega_{1,10}(X) &= \frac{\epsilon_1(X)e_{10}(X) + \epsilon_1(X)e_{11}(X)}{\epsilon_1(X)e_{10}(X) + e_{11}(X)}, & \omega_{0,10}(X) &= \frac{\epsilon_0(X)e_{10}(X) + \epsilon_0(X)e_{00}(X)}{\epsilon_0(X)e_{10}(X) + e_{00}(X)},\\
\omega_{0,00}(X) &= \frac{e_{10}(X) + e_{00}(X)}{\epsilon_0(X)e_{10}(X) + e_{00}(X)}, & \omega_{1,11}(X) &= \frac{e_{10}(X) + e_{11}(X)}{\epsilon_1(X)e_{10}(X) + e_{11}(X)}.
\end{aligned}$$

The following identification formulas hold for the PCEs.

(a) Based on the treatment probability and principal score,

$$\begin{aligned}
\tau_{10} &= \mathbb{E}\left\{\frac{\omega_{1,10}(X)e_{10}(X)}{p_1 - p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{\omega_{0,10}(X)e_{10}(X)}{p_1 - p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\},\\
\tau_{00} &= \mathbb{E}\left\{\frac{1-S}{1-p_1}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{\omega_{0,00}(X)e_{00}(X)}{1-p_1}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\},
\end{aligned}$$

$$\tau_{11} = \mathbb{E}\left\{\frac{\omega_{1,11}(X)e_{11}(X)}{p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{S}{p_0}\frac{1-Z}{1-\pi(X)}Y\right\}.$$

(b) Based on the treatment probability and outcome mean,

$$\begin{aligned}
\tau_{10} &= \mathbb{E}\left[\left\{\frac{SZ}{\pi(X)} - \frac{\mathrm{s}(1-Z)}{1-\pi(X)}\right\}\frac{\omega_{1,10}(X)\mu_{11}(X) - \omega_{0,10}(X)\mu_{00}(X)}{p_1 - p_0}\right], \\
\tau_{00} &= \mathbb{E}\left[\left\{1 - \frac{SZ}{\pi(X)}\right\}\frac{\mu_{10}(X) - \omega_{0,00}(X)\mu_{00}(X)}{p_1 - p_0}\right], \\
\tau_{11} &= \mathbb{E}\left[\frac{S(1-Z)}{1-\pi(X)}\frac{\omega_{1,11}(X)\mu_{11}(X) - \mu_{01}(X)}{p_1 - p_0}\right].
\end{aligned}$$

(c) Based on the principal score and outcome mean,

$$\begin{aligned}
\tau_{10} &= \mathbb{E}\left[\frac{p_1(X) - p_0(X)}{p_1 - p_0}\{\omega_{1,10}(X)\mu_{11}(X) - \omega_{0,10}(X)\mu_{00}(X)\}\right], \\
\tau_{00} &= \mathbb{E}\left[\frac{1 - p_1(X)}{1 - p_1}\{\mu_{10}(X) - \omega_{0,00}(X)\mu_{00}(X)\}\right], \\
\tau_{11} &= \mathbb{E}\left[\frac{p_0(X)}{p_0}\{\omega_{1,11}(X)\mu_{11}(X) - \mu_{01}(X)\}\right].
\end{aligned}$$

Theorem S5 is analogous to Theorem 1. These formulas motivate the estimators for the PCEs by replacing the components with their empirical versions. Similar to Section 4.1, we can derive the EIFs for the PCEs under (S5) as a guidance to propose more principled estimators.

Theorem S6 *Suppose τ_u's are identified in Theorem S5. The EIF for τ_{10} is $\phi'_{10} = \{\phi'_{1,10} - \phi'_{0,10} - \tau_{10}(\psi_{S_1} - \psi_{S_0})\}/(p_1 - p_0)$, where*

$$\begin{aligned}
\phi'_{1,10} &= \frac{\omega_{1,10}(X)e_{10}(X)}{p_1(X)}\psi_{Y_1S_1} - \frac{\omega^2_{1,10}(X)\mu_{11}(X)}{\epsilon_1(X)}\left\{\psi_{S_0} - \frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\}, \\
\phi'_{0,10} &= \frac{\omega_{0,10}(X)e_{10}(X)}{1 - p_0(X)}\psi_{Y_0(1-S_0)} - \frac{\omega^2_{0,10}(X)\mu_{00}(X)}{\epsilon_0(X)}\left\{\psi_{1-S_1} - \frac{1 - p_1(X)}{1 - p_0(X)}\psi_{1-S_0}\right\}.
\end{aligned}$$

The EIF for τ_{00} is $\phi'_{00} = \left(\phi'_{1,00} - \phi'_{0,00} - \tau_{00}\psi_{1-S_1}\right)/(1 - p_1)$, where $\phi'_{1,00} = \phi_{1,00}$ and

$$\phi'_{0,00} = \frac{\omega_{0,00}(X)e_{00}(X)}{1 - p_0(X)}\psi_{Y_0(1-S_0)} + \frac{\omega^2_{0,00}(X)\mu_{00}(X)}{\epsilon_0(X)}\left\{\psi_{1-S_1} - \frac{1 - p_1(X)}{1 - p_0(X)}\psi_{1-S_0}\right\}.$$

The EIF for τ_{11} is $\phi'_{11} = \left(\phi'_{1,11} - \phi'_{0,11} - \tau_{11}\psi_{S_0}\right)/p_0$, where $\phi'_{0,11} = \phi_{0,11}$ and

$$\phi'_{1,11} = \frac{\omega_{1,11}(X)e_{11}(X)}{p_1(X)}\psi_{Y_1S_1} + \frac{\omega^2_{1,11}(X)\mu_{11}(X)}{\epsilon_1(X)}\left\{\psi_{S_0} - \frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\}.$$

We can construct estimators for the PCEs by solving the estimating equation $\mathbb{P}_n\phi'_u = 0$ with the components replaced by their empirical versions:

$$\widehat{\tau}'_{10} = \frac{\mathbb{P}_n(\widehat{\phi}'_{1,10} - \widehat{\phi}'_{0,10})}{\mathbb{P}_n(\widehat{\psi}_{S_1} - \widehat{\psi}_{S_0})}, \quad \widehat{\tau}'_{00} = \frac{\mathbb{P}_n(\widehat{\phi}'_{1,00} - \widehat{\phi}'_{0,00})}{\mathbb{P}_n(1 - \widehat{\psi}_{S_1})}, \quad \widehat{\tau}'_{11} = \frac{\mathbb{P}_n(\widehat{\phi}'_{1,11} - \widehat{\phi}'_{0,11})}{\mathbb{P}_n(\widehat{\psi}_{S_0})}.$$

When $\epsilon_1(X) = \epsilon_0(X) = 1$, these estimators reduce to $\widehat{\tau}_u$ given in Section 4.1. Unlike $\widehat{\tau}_u$, they are not triply robust but only doubly robust, similar to Tchetgen Tchetgen and Shpitser (2012)'s sensitivity analysis for the natural direct and indirect effects.

Theorem S7 *Suppose Assumptions 1, 2, and* (S5) *hold with known* $\epsilon_1(X) \neq 0$ *and* $\epsilon_0(X) \neq 0$. *Suppose further that* $\delta < \{\pi(x;\alpha^*), \pi(x;\hat{\alpha})\} < 1-\delta$, *and* $\{p_1(x;\gamma^*), p_1(x;\widehat{\gamma}), 1-p_0(x;\gamma^*), 1-p_0(x;\widehat{\gamma})\} > \delta$ *for some* $\delta \in (0,1)$ *and all* x *in the support of* X. *The estimator* $\widehat{\tau}'_u$ $(u = 10, 00, 11)$ *is doubly robust in the sense that it is consistent for* τ_u *under* $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{ps+om}}$. *Moreover,* $\widehat{\tau}'_u$ *has the influence function* ϕ_u *and therefore achieves the semiparametric efficiency bound under* $\mathcal{M}_{\text{tp+ps+om}}$.

Our sensitivity analysis estimators require a correct principal score model, weakening the triple robustness to double robustness.

S4.2 Sensitivity analysis for principal ignorability: examples

We first re-visit the example in Section 6.1 to assess the robustness of the conclusions to the violation of principal ignorability. For the ease of presentation, we assume the sensitivity parameters not dependent on X, i.e, $\epsilon_0(X) = \epsilon_0$ and $\epsilon_1(X) = \epsilon_1$, and vary them from 0.75 to 1.25. The upper panel of Figure S1 displays the contour plots of the estimated τ_{10}, τ_{00}, and τ_{11} with different values of (ϵ_0, ϵ_1). When principal ignorability holds ($\epsilon_0 = 1, \epsilon_1 = 1$), $\widehat{\tau}_{10}$ is positive. However, the estimate varies from negative to positive with small changes of (ϵ_0, ϵ_1) from 0.75 to 1.25. This implies that the result is sensitive to the violation of principal ignorability. Similar discussions apply to $\widehat{\tau}_{00}$ and $\widehat{\tau}_{11}$.

We then re-visit the example in Section 6.2. The lower panel of Figure S1 displays the contour plots of the estimated parameters. Unlike the example in Section 6.1, the result is not very sensitive to the violation of principal ignorability. In particular, $\widehat{\tau}_{10}$ changes the sign only when ϵ_1 becomes close to 1.25, $\widehat{\tau}_{00}$ remains positive in a wide range of ϵ_0, and $\widehat{\tau}_{11}$ remains positive for all values of ϵ_1.

S4.3 Sensitivity analysis for monotonicity

Without monotonicity, we have an additional stratum $U = 01$. In this case, the proportions of principal strata are not identifiable without further assumptions. Define the sensitivity parameter as the ratio between the proportion of stratum 01 and stratum 10 conditional on the covariates,

$$\xi(X) \;=\; \frac{\Pr(U = 01 \mid X)}{\Pr(U = 10 \mid X)}. \tag{S6}$$

It characterizes the deviation from monotonicity. If $\xi(X) = 0$, then monotonicity holds. Ding and Lu (2017) discuss a special case with $\xi(X)$ being a constant.

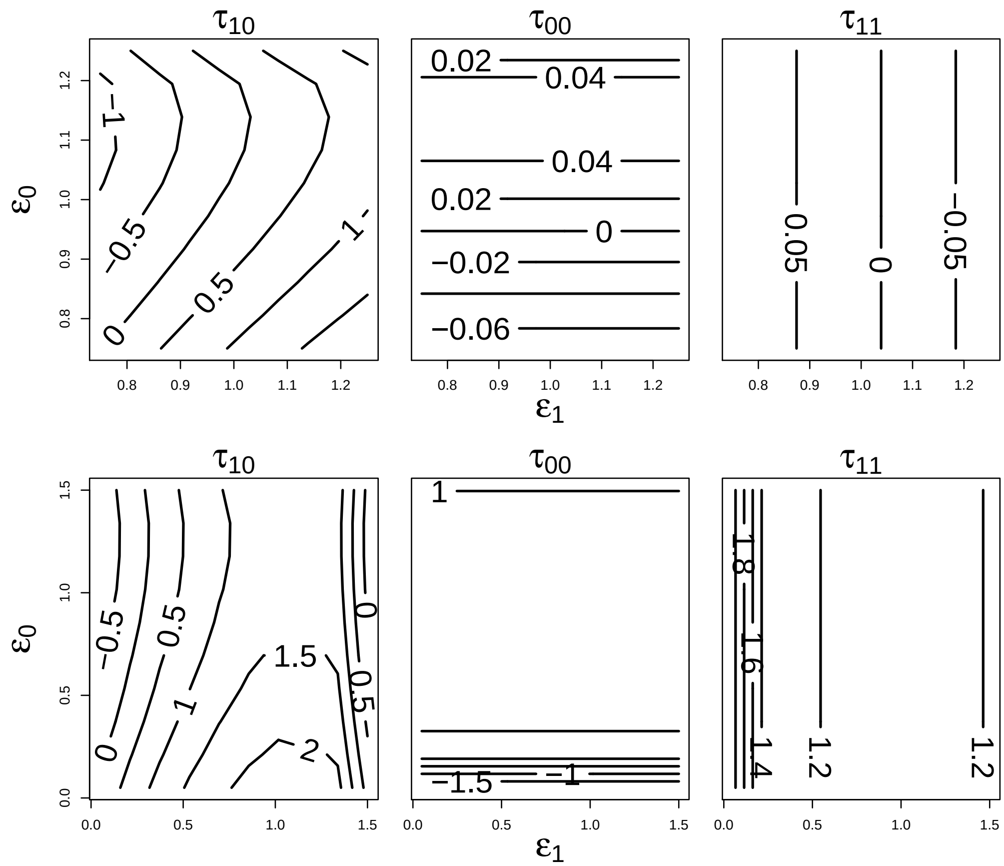

Figure S1: Sensitivity analysis for Sections 6.1 (top) and 6.2 (bottom).

For a fixed sensitivity parameter $\xi(X) \neq 1$, we can identify the principal scores by

$$\begin{array}{ll} e_{\xi,10}(X) = \frac{p_1(X)-p_0(X)}{1-\xi(X)}, & e_{\xi,00}(X) = 1 - p_0(X) - \frac{p_1(X)-p_0(X)}{1-\xi(X)}, \\ e_{\xi,11}(X) = p_1(X) - \frac{p_1(X)-p_0(X)}{1-\xi(X)}, & e_{\xi,01}(X) = \frac{\xi(X)\{p_1(X)-p_0(X)\}}{1-\xi(X)} \end{array} \tag{S7}$$

and the proportions of principal strata by $e_{\xi,u} = \mathbb{E}\{e_{\xi,u}(X)\}$ for all u. When $\xi(X) = 1$, strata 10 and 01 have equal proportions. This corresponds to the boundary case when the treatment has zero average causal effect on S. We rule out this boundary case because the principal scores are not identifiable.

Without monotonicity, we need to extend principal ignorability to include stratum 01.

Assumption S3 (Principal ignorability without monotonicity) $\mathbb{E}(Y_1 \mid U = s_1 1, X) = \mathbb{E}(Y_1 \mid U = s_1 0, X)$ *for* $s_1 = 0, 1$ *and* $\mathbb{E}(Y_0 \mid U = 1 s_0, X) = E(Y_0 \mid U = 0 s_0, X)$ *for* $s_0 = 0, 1$.

The following theorem establishes the nonparametric identification of the PCEs with a known $\xi(X)$.

Theorem S8 *Under Assumptions 1, S3, and* (S6) *with a known* $\xi(X) \neq 1$, $e_{\xi,u} > 0$ *for* $u = 10, 00, 11$, *and* $0 < \pi(x) < 1$ *for all* x *in the support of* X. *The following identification formulas hold for the PCEs.*

(a) Based on the treatment probability and principal score,

$$\begin{aligned}
\tau_{10} &= \mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\}, \\
\tau_{00} &= \mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\}, \\
\tau_{11} &= \mathbb{E}\left\{\frac{e_{\xi,11}(X)}{e_{\xi,11}}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{\xi,11}(X)}{e_{\xi,11}}\frac{S}{p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\}, \\
\tau_{01} &= \mathbb{E}\left\{\frac{e_{\xi,01}(X)}{e_{\xi,01}}\frac{1-S}{1-p_1(X)}\frac{Z}{\pi(X)}Y\right\} - \mathbb{E}\left\{\frac{e_{\xi,01}(X)}{e_{\xi,01}}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\}.
\end{aligned}$$

(b) Based on the treatment probability and outcome mean,

$$\begin{aligned}
\tau_{10} &= \mathbb{E}\left[\left\{\frac{SZ}{\{1-\xi(X)\}\pi(X)} - \frac{S(1-Z)}{\{1-\xi(X)\}\{1-\pi(X)\}}\right\}\frac{\mu_{11}(X)-\mu_{00}(X)}{e_{\xi,10}}\right], \\
\tau_{00} &= \mathbb{E}\left[\left\{1 - \frac{SZ}{\{1-\xi(X)\}\pi(X)} - \frac{\xi(X)S(1-Z)}{\{1-\xi(X)\}\{1-\pi(X)\}}\right\}\frac{\mu_{10}(X)-\mu_{00}(X)}{e_{\xi,00}}\right], \\
\tau_{11} &= \mathbb{E}\left[\left\{-\frac{\xi(X)SZ}{\{1-\xi(X)\}\pi(X)} + \frac{S(1-Z)}{\{1-\xi(X)\}\{1-\pi(X)\}}\right\}\frac{\mu_{11}(X)-\mu_{01}(X)}{e_{\xi,11}}\right], \\
\tau_{01} &= \mathbb{E}\left[\left\{\frac{\xi(X)SZ}{\{1-\xi(X)\}\pi(X)} - \frac{\xi(X)S(1-Z)}{\{1-\xi(X)\}\{1-\pi(X)\}}\right\}\frac{\mu_{10}(X)-\mu_{01}(X)}{e_{\xi,01}}\right].
\end{aligned}$$

(c) Based on the principal score and outcome mean,

$$\begin{aligned}
\tau_{10} &= \mathbb{E}\left[\frac{e_{\xi,10}(X)}{e_{\xi,10}}\{\mu_{11}(X)-\mu_{00}(X)\}\right], \\
\tau_{00} &= \mathbb{E}\left[\frac{e_{\xi,00}(X)}{e_{\xi,00}}\{\mu_{10}(X)-\mu_{00}(X)\}\right], \\
\tau_{11} &= \mathbb{E}\left[\frac{e_{\xi,11}(X)}{e_{\xi,11}}\{\mu_{11}(X)-\mu_{01}(X)\}\right], \\
\tau_{01} &= \mathbb{E}\left[\frac{e_{\xi,01}(X)}{e_{\xi,01}}\{\mu_{10}(X)-\mu_{01}(X)\}\right].
\end{aligned}$$

Theorem S8 is analogous to Theorem 1. These formulas motivate the estimators for the PCEs by replacing the components with their empirical versions. Similar to Section 4.1, we can derive the EIFs for the PCEs under (S6) as guidance to propose more principled estimators. For simplicity, we only give the result for τ_{10}.

Theorem S9 *Suppose* τ_u*'s are identified in Theorem S8. The EIF for* τ_{10} *is*

$$\phi^*_{10} = \frac{1}{e_{\xi,10}}\left\{(\phi^*_{1,10} - \phi^*_{0,10}) - \frac{\tau_{10}(\psi_{S_1} - \psi_{S_0})}{1-\xi(X)}\right\},$$

where

$$
\begin{aligned}
\phi^*_{1,10} &= \frac{p_1(X)-p_0(X)}{\{1-\xi(X)\}p_1(X)}\psi_{Y_1S_1} - \frac{\mu_{11}(X)}{1-\xi(X)}\left\{\psi_{S_0} - \frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\}, \\
\phi^*_{0,10} &= \frac{p_1(X)-p_0(X)}{\{1-\xi(X)\}\{1-p_0(X)\}}\psi_{Y_0(1-S_0)} - \frac{\mu_{00}(X)}{1-\xi(X)}\left\{\psi_{1-S_1} - \frac{1-p_1(X)}{1-p_0(X)}\psi_{1-S_0}\right\}.
\end{aligned}
$$

We can construct estimators for the PCEs by solving the estimating equation $\mathbb{P}_n\phi^*_u = 0$ with the components replaced by their empirical versions, e.g.,

$$
\widehat{\tau}'_{10} = \frac{\mathbb{P}_n(\widehat{\phi}^*_{1,10} - \widehat{\phi}^*_{0,10})}{\mathbb{P}_n(\widehat{\psi}_{S_1} - \widehat{\psi}_{S_0})}.
$$

When $\xi(X) = 0$, these estimators reduce to $\widehat{\tau}_u$ given in Section 4.1. The following theorem shows the triple robustness and local efficiency of the estimators constructed based on the EIFs.

Theorem S10 *Suppose Assumptions 1, S3, and* (S6) *hold with a known* $\xi(X) \neq 1$, $\delta < \{\pi(x;\alpha^*), \pi(x;\hat{\alpha})\} < 1-\delta$, *and* $\{p_1(x;\gamma^*), p_1(x;\widehat{\gamma}), 1-p_0(x;\gamma^*), 1-p_0(x;\widehat{\gamma})\} > \delta$ *for some* $\delta \in (0,1)$ *and all* x *in the support of* X*. The estimator* $\widehat{\tau}^*_u$ $(u = 10, 00, 11, 01)$ *is triply robust in the sense that it is consistent for* τ_u *under* $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$*. Moreover,* $\widehat{\tau}^*_u$ *has the influence function* ϕ^*_u *and therefore achieves the semiparametric efficiency bound under* $\mathcal{M}_{\text{tp+ps+om}}$*.*

S5 Proof of the identification results

Throughout the proofs, we will use $f(\cdot)$ to denote the probability density functions for continuous random variables and the probability mass functions for discrete random variables. We will use the law of total expectation repeatedly and will mark the steps with LOTE. We will use the following lemma repeatedly, the proof of which is straightforward.

Lemma S1 *Let* X *and* Y *be two random variables with densities* $f_1(x)$ *and* $f_2(y)$*. For any function* $h(\cdot)$ *with* $\mathbb{E}\{h(X)\} < \infty$*, we have*

$$
\mathbb{E}\{h(X)\} = \mathbb{E}\left\{\frac{f_1(Y)}{f_2(Y)}h(Y)\right\}.
$$

S5.1 Proof of Theorem S1

It follows from the LOTP and Lemma S1 by conditioning on X. We omit the details due to the length of the supplementary material. The earliest ArXiv version of our paper contains the details.

S5.2 Proof of Theorem 1

We prove only the identification formulas for τ_{10} and omit the similar proofs of the identification formulas for τ_{00} and τ_{11}. We have

$$\begin{aligned}
\tau_{10} &= \mathbb{E}(Y_1 \mid U=10) - \mathbb{E}(Y_0 \mid U=10) \\
&= \mathbb{E}\{\mathbb{E}(Y_1 \mid U=10, X) \mid U=10\} - \mathbb{E}\{\mathbb{E}(Y_0 \mid U=10, X) \mid U=10\} \\
&= \mathbb{E}\{\mathbb{E}(Y_1 \mid U=10 \text{ or } 11, X) \mid U=10\} \\
&\quad -\mathbb{E}\{\mathbb{E}(Y_0 \mid U=10 \text{ or } 00, X) \mid U=10\} \quad \text{(Assumption 3)} \\
&= \mathbb{E}\{\mathbb{E}(Y \mid Z=1, U=10 \text{ or } 11, X) \mid U=10\} \\
&\quad -\mathbb{E}\{\mathbb{E}(Y \mid Z=0, U=10 \text{ or } 00, X) \mid U=10\} \quad \text{(Assumption 1)} \\
&= \mathbb{E}\{\mathbb{E}(Y \mid Z=1, S=1, X) \mid U=10\} - \mathbb{E}\{\mathbb{E}(Y \mid Z=0, S=0, X) \mid U=10\} \\
&= \mathbb{E}\{\mu_{11}(X) \mid U=10\} - \mathbb{E}\{\mu_{00}(X) \mid U=10\}. \qquad \text{(S8)}
\end{aligned}$$

Theorem S1 ensures that for any $h(X)$,

$$\begin{aligned}
\mathbb{E}\{h(X) \mid U=10\} &= \mathbb{E}\left[\left\{\frac{SZ}{\pi(X)} - \frac{S(1-Z)}{1-\pi(X)}\right\}\frac{h(X)}{p_1-p_0}\right], \qquad \text{(S9)} \\
&= \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}h(X)\right\}. \qquad \text{(S10)}
\end{aligned}$$

Applying (S9) to the two terms in (S8), we obtain

$$\tau_{10} = \mathbb{E}\left[\left\{\frac{SZ}{\pi(X)} - \frac{S(1-Z)}{1-\pi(X)}\right\}\frac{\mu_{11}(X)-\mu_{00}(X)}{p_1-p_0}\right],$$

which is the identification formula in Theorem 1(b). Applying (S10) to the two terms in (S8), we obtain

$$\tau_{10} = \mathbb{E}\left[\frac{p_1(X)-p_0(X)}{p_1-p_0}\{\mu_{11}(X)-\mu_{00}(X)\}\right], \qquad \text{(S11)}$$

which is the identification formula in Theorem 1(c).

We finally prove the identification formula in Theorem 1(a). We have,

$$\begin{aligned}
\mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} &= \mathbb{E}\left[\mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y \mid X\right\}\right] \quad \text{(LOTE)} \\
&= \mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{\mathbb{P}(Z=1, S=1 \mid X)}{p_1(X)\pi(X)}\mu_{11}(X)\right\} \\
&= \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}\mu_{11}(X)\right\}, \\
\mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\} &= \mathbb{E}\left[\mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y \mid X\right\}\right] \quad \text{(LOTE)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left\{\frac{e_{10}(X)}{p_1-p_0}\frac{\mathbb{P}(Z=0,S=0\mid X)}{\{1-p_0(X)\}\{1-\pi(X)\}}\mu_{00}(X)\right\} \\
&= \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}\mu_{00}(X)\right\},
\end{aligned}
$$

which, coupled with (S11), imply the identification formula in Theorem 1(a). □

S5.3 Proof of Theorem S5

We prove only the identification formulas for τ_{10} and omit the similar proofs of the identification formulas for τ_{00} and τ_{11}. We have,

$$
\begin{aligned}
\mu_{11}(X) &= \mathbb{E}(Y_1 \mid Z=1,S=1,U=10,X)\mathbb{P}(U=10\mid Z=1,S=1,X) \\
&\quad +\mathbb{E}(Y_1 \mid Z=1,S=1,U=11,X)\mathbb{P}(U=11\mid Z=1,S=1,X) \quad \text{(LOTE)} \\
&= \mathbb{E}(Y_1 \mid U=10,X)\mathbb{P}(U=10\mid Z=1,S=1,X) \\
&\quad +\mathbb{E}(Y_1 \mid U=11,X)\mathbb{P}(U=11\mid Z=1,S=1,X) \quad \text{(Assumption 3)} \\
&= \mathbb{E}(Y_1 \mid U=10,X)\frac{e_{10}(X)}{e_{10}(X)+e_{11}(X)} + \mathbb{E}(Y_1 \mid U=11,X)\frac{e_{11}(X)}{e_{10}(X)+e_{11}(X)} \\
&= \mathbb{E}(Y_1 \mid U=10,X)\frac{\epsilon_1(X)e_{10}(X)+e_{11}(X)}{\epsilon_1(X)\{e_{10}(X)+e_{11}(X)\}}. \quad \text{(by (S5))}
\end{aligned}
$$

Therefore,

$$
\mathbb{E}(Y_1 \mid U=10,X) = \frac{\epsilon_1(X)\{e_{10}(X)+e_{11}(X)\}}{\epsilon_1(X)e_{10}(X)+e_{11}(X)}\mu_{11}(X) = \omega_{1,10}(X)\mu_{11}(X). \tag{S12}
$$

Similarly,

$$
\mathbb{E}(Y_0 \mid U=10,X) = \frac{\epsilon_0(X)\{e_{10}(X)+e_{00}(X)\}}{\epsilon_0(X)e_{10}(X)+e_{00}(X)}\mu_{00}(X) = \omega_{0,10}(X)\mu_{00}(X).
$$

The above two formulas imply

$$
\begin{aligned}
\tau_{10} &= \mathbb{E}\{\mathbb{E}(Y_1 \mid U=10,X)\mid U=10\} - \mathbb{E}\{\mathbb{E}(Y_0 \mid U=10,X)\mid U=10\} \\
&= \mathbb{E}\left\{\omega_{1,10}(X)\mu_{11}(X)\mid U=10\right\} - \mathbb{E}\left\{\omega_{0,10}(X)\mu_{00}(X)\mid U=10\right\}.
\end{aligned} \tag{S13}
$$

Applying (S9) to the two terms in (S13), we obtain

$$
\tau_{10} = \mathbb{E}\left[\left\{\frac{SZ}{\pi(X)} - \frac{S(1-Z)}{1-\pi(X)}\right\}\frac{\omega_{1,10}(X)\mu_{11}(X)-\omega_{0,10}(X)\mu_{00}(X)}{p_1-p_0}\right], \tag{S14}
$$

which is the identification formula in Theorem S5(b). Applying (S10) to the two terms in (S13), we obtain

$$
\tau_{10} = \mathbb{E}\left[\frac{p_1(X)-p_0(X)}{p_1-p_0}\{\omega_{1,10}(X)\mu_{11}(X)-\omega_{0,10}(X)\mu_{00}(X)\}\right], \tag{S15}
$$

which is the identification formula in Theorem S5(c).

We finally prove the identification formula in Theorem S5(a). We have

$$
\begin{aligned}
& \mathbb{E}\left\{\frac{\omega_{1,10}(X)e_{10}(X)}{p_1-p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} \\
= \; & \mathbb{E}\left[\mathbb{E}\left\{\frac{\omega_{1,10}(X)e_{10}(X)}{p_1-p_0}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} \mid X\right] \quad \text{(LOTE)} \\
= \; & \mathbb{E}\left\{\frac{\omega_{1,10}(X)e_{10}(X)}{p_1-p_0}\frac{\mathbb{P}(Z=1, S=1 \mid X)}{p_1(X)\pi(X)}\mu_{11}(X)\right\} \\
= \; & \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}\omega_{1,10}(X)\mu_{11}(X)\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
& \mathbb{E}\left\{\frac{\omega_{0,10}(X)e_{10}(X)}{p_1-p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\} \\
= \; & \mathbb{E}\left[\mathbb{E}\left\{\frac{\omega_{0,10}(X)e_{10}(X)}{p_1-p_0}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y \mid X\right\}\right] \quad \text{(LOTE)} \\
= \; & \mathbb{E}\left\{\frac{\omega_{0,10}(X)e_{10}(X)}{p_1-p_0}\frac{\mathbb{P}(Z=0, S=0 \mid X)}{\{1-p_0(X)\}\{1-\pi(X)\}}\mu_{00}(X)\right\} \\
= \; & \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{p_1-p_0}\omega_{0,10}(X)\mu_{00}(X)\right\},
\end{aligned}
$$

which, coupled with (S15), imply the identification formula in Theorem S5(a). □

S5.4 Proof of Theorem S8

We prove only the identification formulas for τ_{10} and omit the similar proofs of the identification formulas for τ_{00}, τ_{11}, and τ_{01}. Under Assumptions 1 and S3, we have

$$
\mathbb{E}(Y_1 \mid U=10, X) \;=\; \mathbb{E}(Y_1 \mid S_1=1, X) = \mu_{11}(X).
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}(Y_1 \mid U=10) \;&=\; \mathbb{E}\{\mathbb{E}(Y_1 \mid U=10, X) \mid U=10\} \\
&=\; \mathbb{E}\{\mu_{11}(X) \mid U=10\} \\
&=\; \mathbb{E}\left\{\frac{\Pr(U=10 \mid X)}{\Pr(U=10)}\mu_{11}(X)\right\} \quad \text{(Lemma S1)} \\
&=\; \mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\mu_{11}(X)\right\}.
\end{aligned}
$$

Similarly,

$$
\mathbb{E}(Y_0 \mid U=10) \;=\; \mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\mu_{00}(X)\right\}.
$$

The above two formulas imply

$$\tau_{10} \quad = \quad \mathbb{E}\left[\frac{e_{\xi,10}(X)}{e_{\xi,10}}\{\mu_{11}(X)-\mu_{00}(X)\}\right], \tag{S16}$$

which is the identification formula in Theorem S8(c). Applying Theorem S1(a) to (S16), we have

$$\begin{aligned}\tau \quad &= \quad \mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\frac{\mu_{11}(X)-\mu_{00}(X)}{e_{\xi,10}}\right\} \\ &= \quad \mathbb{E}\left[\left\{\frac{SZ}{\pi(X)}-\frac{S(1-Z)}{1-\pi(X)}\right\}\frac{\mu_{11}(X)-\mu_{00}(X)}{\{1-\xi(X)\}e_{\xi,10}}\right],\end{aligned}$$

which is the identification formula in Theorem S8(b).

We finally prove the identification formula in Theorem S8(a). We have

$$\begin{aligned}&\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} \\ = \quad &\mathbb{E}\left[\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{S}{p_1(X)}\frac{Z}{\pi(X)}Y\right\} \mid X\right] \quad \text{(LOTE)} \\ = \quad &\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{\mathbb{P}(Z=1,S=1\mid X)}{p_1(X)\pi(X)}\mu_{11}(X)\right\} \\ = \quad &\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\mu_{11}(X)\right\}\end{aligned} \tag{S17}$$

and

$$\begin{aligned}&\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{1-S}{1-p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\} \\ = \quad &\mathbb{E}\left[\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{1-S}{p_0(X)}\frac{1-Z}{1-\pi(X)}Y\right\} \mid X\right] \quad \text{(LOTE)} \\ = \quad &\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\frac{\mathbb{P}(Z=0,S=1\mid X)}{\{1-p_0(X)\}\{1-\pi(X)\}}\mu_{00}(X)\right\} \\ = \quad &\mathbb{E}\left\{\frac{e_{\xi,10}(X)}{e_{\xi,10}}\mu_{00}(X)\right\},\end{aligned} \tag{S18}$$

which, coupled with (S16), imply the identification formula in Theorem S8(a). □

S6 Proof of the EIFs

In this section, we prove Theorems 2, S2, S3, S4, S6, and S9.

S6.1 Proof of Theorem 2

S6.1.1 Preliminaries

We will use the semiparametric theory in Bickel et al. (1993) to derive the EIFs. Let $V=(X,Z,S,Y)$ be the vector of all observed variables with the likelihood factorized as

$$f(V) \quad = \quad f(X)f(Z\mid X)f(S\mid Z,X)f(Y\mid Z,S,X). \tag{S19}$$

For $z = 0, 1$ and $u = 10, 00, 11$, let $\mu_{z,u}$ be the identification of $\mathbb{E}(Y_z \mid U = u)$ given in Theorem 1. Then, we have $\tau_u = \mu_{1,u} - \mu_{0,u}$. To derive the EIFs, we consider a one-dimensional parametric submodel, $f_\theta(V)$, which contains the true model $f(V)$ at $\theta = 0$, i.e., $f_\theta(V)|_{\theta=0} = f(V)$. We use θ in the subscript to denote the quantity with respect to the submodel, e.g., $\mu_{z,u,\theta}$ is the value of $\mu_{z,u}$ in the submodel. We use dot to denote the partial derivative with respect to θ, e.g., $\dot{\mu}_{z,u,\theta} = \partial\mu_{z,u,\theta}/\partial\theta$, and use $\mathrm{s}_\theta(\cdot)$ to denote the score function of the submodel. From (S19), the score function under the submodel of can be decomposed as

$$\mathrm{s}_\theta(V) = \mathrm{s}_\theta(X) + \mathrm{s}_\theta(Z \mid X) + \mathrm{s}_\theta(S \mid Z, X) + \mathrm{s}_\theta(Y \mid S, Z, X),$$

where $\mathrm{s}_\theta(X) = \partial \log f_\theta(X)/\partial\theta$, $\mathrm{s}_\theta(Z \mid X) = \partial \log f_\theta(Z \mid X)/\partial\theta$, $\mathrm{s}_\theta(S \mid Z, X) = \partial \log f_\theta(S \mid Z, X)/\partial\theta$, and $\mathrm{s}_\theta(Y \mid S, Z, X) = \partial \log f_\theta(Y \mid S, Z, X)/\partial\theta$ are the score functions corresponding to the four components of the likelihood. Analogous to $f_\theta(V)|_{\theta=0} = f(V)$, we write $\mathrm{s}_\theta(\cdot)|_{\theta=0}$ as $\mathrm{s}(\cdot)$, which is the score function evaluated at the true parameter under the one-dimensional submodel.

From the semiparametric theory, the tangent space

$$\Lambda = H_1 \oplus H_2 \oplus H_3 \oplus H_4 \tag{S20}$$

is the direct sum of

$$\begin{aligned}
H_1 &= \{h(X) : \mathbb{E}\{h(X)\} = 0\}, \\
H_2 &= \{h(Z, X) : \mathbb{E}\{h(Z, X) \mid X\} = 0\}, \\
H_3 &= \{h(S, Z, X) : \mathbb{E}\{h(S, Z, X) \mid Z, X\} = 0\}, \\
H_4 &= \{h(Y, Z, S, X) : \mathbb{E}\{h(Y, Z, S, X) \mid Z, S, X\} = 0\},
\end{aligned}$$

where H_1, H_2, H_3, and H_4 are orthogonal to each other. The EIF for $\mu_{z,u}$, denoted by $\varphi_{z,u}(V) \in \Lambda$, must satisfy

$$\dot{\mu}_{z,u,\theta}|_{\theta=0} = \mathbb{E}\{\varphi_{z,u}(V)\mathrm{s}(V)\}.$$

We will derive the EIFs by calculating $\dot{\mu}_{z,u,\theta}|_{\theta=0}$. To simplify the proof, we introduce some lemmas.

Lemma S2 *Consider a ratio-type parameter $R = N/D$. If $\dot{N}_\theta|_{\theta=0} = \mathbb{E}\{\varphi_N(V)\mathrm{s}(V)\}$ and $\dot{D}_\theta|_{\theta=0} = \mathbb{E}\{\varphi_D(V)\mathrm{s}(V)\}$, then $\dot{R}_\theta|_{\theta=0} = \mathbb{E}\{\varphi_R(V)\mathrm{s}(V)\}$ where*

$$\varphi_R(V) = \frac{1}{D}\varphi_N(V) - \frac{R}{D}\varphi_D(V). \tag{S21}$$

In particular, if $\varphi_N(V)$ and $\varphi_D(V)$ are the EIFs for N and D, then $\varphi_R(V)$ is the EIF for R.

Proof of Lemma S2. Let R_θ, N_θ, and D_θ denote the quantities R, N, and D evaluated with respect to the parametric submodel $f_\theta(V)$. By the chain rule, we have

$$
\begin{aligned}
\dot{R}_\theta\Big|_{\theta=0} &= \frac{\dot{N}_\theta}{D}\Bigg|_{\theta=0} - R_\theta\,\frac{\dot{D}_\theta}{D}\Bigg|_{\theta=0} \\
&= \frac{1}{D}\mathbb{E}\{\varphi_N(V)\mathrm{s}(V)\} - \frac{R}{D}\mathbb{E}\{\varphi_D(V)\mathrm{s}(V)\} \\
&= \mathbb{E}\left[\left\{\frac{1}{D}\varphi_N(V) - \frac{R}{D}\varphi_D(V)\right\}\mathrm{s}(V)\right],
\end{aligned}
$$

which yields (S21). □

Lemma S3 *For any $h(V)$ that does not depend on θ, $\partial\mathbb{E}_\theta\{h(V)\}/\partial\theta\mid_{\theta=0}=\mathbb{E}\{h(V)\mathrm{s}(V)\}$.*

The proof is straightforward and thus omitted.

Lemma S4 *Define $\mu_{0f}(X)=\mathbb{E}\{f(Y,S,X)\mid Z=0,X\}$ and $\mu_{1f}(X)=\mathbb{E}\{f(Y,S,X)\mid Z=1,X\}$ for any $f(Y,S,X)$. We have*

$$
\begin{aligned}
\dot{\mu}_{0f,\theta}(X)|_{\theta=0} &= \mathbb{E}\left[\{\psi_{f(Y_0,S_0,X)} - \mu_{0f}(X)\}\mathrm{s}(Y,S\mid Z,X)\mid X\right], \\
\dot{\mu}_{1f,\theta}(X)|_{\theta=0} &= \mathbb{E}\left[\{\psi_{f(Y_1,S_1,X)} - \mu_{1f}(X)\}\mathrm{s}(Y,S\mid Z,X)\mid X\right].
\end{aligned}
$$

As a special case, for $p_0(X)=\mathbb{E}(S\mid Z=0,X)$ and $p_1(X)=\mathbb{E}(S\mid Z=1,X)$, we have

$$
\begin{aligned}
\dot{p}_{0,\theta}(X)\mid_{\theta=0} &= \mathbb{E}\left[\{\psi_{S_0} - p_0(X)\}\mathrm{s}(S\mid Z,X)\mid X\right], \\
\dot{p}_{1,\theta}(X)\mid_{\theta=0} &= \mathbb{E}\left[\{\psi_{S_1} - p_1(X)\}\mathrm{s}(S\mid Z,X)\mid X\right].
\end{aligned}
$$

Proof of Lemma S4. We first prove the general result:

$$
\begin{aligned}
\dot{\mu}_{0f,\theta}(X)|_{\theta=0} &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta\{f(Y,S,X)\mid Z=0,X\}\mid_{\theta=0} \\
&= \mathbb{E}\left\{f(Y,S,X)\times\mathrm{s}(Y,S\mid Z=0,X)\mid Z=0,X\right\} \quad \text{(Lemma S3)} \\
&= \mathbb{E}\left[\{f(Y,S,X)-\mu_{0f}(X)\}\mathrm{s}(Y,S\mid Z=0,X)\mid Z=0,X\right] \\
&= \mathbb{E}\left[\frac{(1-Z)\{f(Y,S,X)-\mu_{0f}(X)\}}{1-\pi(X)}\mathrm{s}(Y,S\mid Z,X)\mid X\right] \\
&= \mathbb{E}\left[\{\psi_{f(Y_0,S_0,X)} - \mu_{0f}(X)\}\mathrm{s}(Y,S\mid Z,X)\mid X\right]
\end{aligned}
$$

where the third equality follows from $\mu_{0f}(X)\mathbb{E}\{\mathrm{s}(Y,S\mid Z=0,X)\mid Z=0,X\}=0$. Similarly, we can prove the result for $\mu_{1f}(X)$.

Choosing $f(Y,S,X)=S$, the results for $p_z(X)$ follow because

$$\mathbb{E}\left[\{\psi_{S_z}-p_z(X)\}\mathrm{s}(Y\mid Z,S,X)\mid X\right]=0.$$

$\square$

Lemma S5 *Define $\mu_{0f}=\mathbb{E}\{\mu_{0f}(X)\}$ and $\mu_{1f}=\mathbb{E}\{\mu_{1f}(X)\}$. We have*

$$\begin{aligned}\dot{\mu}_{0f,\theta}|_{\theta=0} &= \mathbb{E}\left[\{\psi_{f(Y_0,S_0,X)}-\mu_{0f}\}\mathrm{s}(V)\right],\\ \dot{\mu}_{1f,\theta}|_{\theta=0} &= \mathbb{E}\left[\{\psi_{f(Y_1,S_1,X)}-\mu_{1f}\}\mathrm{s}(V)\right].\end{aligned}$$

Moreover, $\psi_{f(Y_0,S_0,X)}-\mu_{0f}$ and $\psi_{f(Y_1,S_1,X)}-\mu_{1f}$ are EIFs for μ_{0f} and μ_{1f}, respectively. As a special case, for $p_0=\mathbb{E}\{p_0(X)\}$ and $p_1=\mathbb{E}\{p_1(X)\}$, we have

$$\begin{aligned}\dot{p}_{0,\theta}\mid_{\theta=0} &= \mathbb{E}\left\{(\psi_{S_0}-p_0)\,\mathrm{s}(V)\right\},\\ \dot{p}_{1,\theta}\mid_{\theta=0} &= \mathbb{E}\left\{(\psi_{S_1}-p_1)\,\mathrm{s}(V)\right\},\end{aligned}$$

and $\psi_{S_0}-p_0$ and $\psi_{S_1}-p_1$ are the EIFs for p_0 and p_1, respectively.

Proof of Lemma S5. We prove the general result:

$$\begin{aligned}\dot{\mu}_{0f,\theta}\mid_{\theta=0} &= \mathbb{E}\{\mu_{0f}(X)\mathrm{s}(V)\}+\mathbb{E}\{\dot{\mu}_{0f,\theta}(X)\mid_{\theta=0}\} \quad \text{(Lemma S3)}\\ &= \mathbb{E}\{\mu_{0f}(X)\mathrm{s}(V)\}+\mathbb{E}\left[\{\psi_{f(Y_0,S_0,X)}-\mu_{0f}(X)\}\mathrm{s}(Y,S\mid Z,X)\right] \quad \text{(Lemma S4)}\\ &= \mathbb{E}\{(\psi_{f(Y_0,S_0,X)}-\mu_{0f})\mathrm{s}(Y,S\mid Z,X)\}\\ &= \mathbb{E}\{(\psi_{f(Y_0,S_0,X)}-\mu_{0f})\mathrm{s}(V)\},\end{aligned}$$

where the last equality follows from $\mathbb{E}\{(\psi_{f(Y_0,S_0,X)}-\mu_{0f})\mathrm{s}(Z\mid X)\}=\mathbb{E}\{(\psi_{f(Y_0,S_0,X)}-\mu_{0f})\mathrm{s}(X)\}=0$. Because $\psi_{f(Y_0,S_0,X)}-\mu_{0f}$ lies in the tangent space, it is the EIF for μ_{0f}. Similarly, we can prove the result for μ_{1f}. The results for p_1 and p_0 follow by taking $f(Y,S,X)=S$. Hahn (1998) gives an alternative proof of the EIFs of p_1 and p_0. $\square$

Lemma S6 *For $\mu_{zs}(X)$, we have*

$$\begin{aligned}\dot{\mu}_{11,\theta}(X)\mid_{\theta=0} &= \mathbb{E}\left\{\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y\mid Z,S,X)\mid X\right\},\\ \dot{\mu}_{01,\theta}(X)\mid_{\theta=0} &= \mathbb{E}\left\{\frac{\psi_{Y_0S_0}-\mu_{01}(X)\psi_{S_0}}{p_0(X)}\mathrm{s}(Y\mid Z,S,X)\mid X\right\},\\ \dot{\mu}_{10,\theta}(X)\mid_{\theta=0} &= \mathbb{E}\left\{\frac{\psi_{Y_1(1-S_1)}-\mu_{10}(X)\psi_{1-S_1}}{1-p_1(X)}\mathrm{s}(Y\mid Z,S,X)\mid X\right\},\\ \dot{\mu}_{00,\theta}(X)\mid_{\theta=0} &= \mathbb{E}\left\{\frac{\psi_{Y_0(1-S_0)}-\mu_{00}(X)\psi_{1-S_0}}{1-p_0(X)}\mathrm{s}(Y\mid Z,S,X)\mid X\right\}.\end{aligned}$$

Proof of Lemma S6. We prove only the result for $\mu_{11,\theta}(X)$ and the proofs for other parameters are similar. A key observation is the ratio representation:

$$\mu_{11}(X) = \mathbb{E}(Y \mid Z = 1, S = 1, X) = \frac{\mathbb{E}(YS \mid Z = 1, X)}{\mathbb{E}(S \mid Z = 1, X)} = \frac{\mathbb{E}(YS \mid Z = 1, X)}{p_1(X)}.$$

From Lemma S4, the numerator satisfies

$$\frac{\partial}{\partial\theta}\mathbb{E}(YS \mid Z = 1, X) \mid_{\theta=0} = \mathbb{E}\left[\{\psi_{Y_1S_1} - p_1(X)\mu_{11}(X)\}\mathrm{s}(Y, S \mid Z, X) \mid X\right],$$

and the denominator satisfies

$$\dot{p}_1(X) \mid_{\theta=0} = \mathbb{E}\left[\{\psi_{S_1} - p_1(X)\}\mathrm{s}(Y, S \mid Z, X) \mid X\right].$$

We can then use Lemma S2 to calculate the path derivative of $\mu_{11,\theta}(X)$ with all distributions conditional on X, yielding

$$\dot{\mu}_{11,\theta}(X) \mid_{\theta=0} = \mathbb{E}\left\{\frac{\psi_{Y_1S_1} - \mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y, S \mid Z, X) \mid X\right\}.$$

The conclusion follows by using $\mathbb{E}[\{\psi_{Y_1S_1} - \mu_{11}(X)\psi_{S_1}\}\mathrm{s}(S \mid Z, X) \mid X] = 0$. □

S6.1.2 EIF for τ_{10}

Below, we derive the EIF for τ_{10}. The EIFs for τ_{00} and τ_{11} can be derived similarly.

First, from Theorem 1(c), we can write $\mu_{1,10} = N/D$, where

$$N = \mathbb{E}\left[\{p_1(X) - p_0(X)\}\mu_{11}(X)\right], \quad D = p_1 - p_0.$$

Lemma S5 implies

$$\varphi_D(V) = (\psi_{S_1} - \psi_{S_0}) - (p_1 - p_0) = (\psi_{S_1} - \psi_{S_0}) - D, \tag{S22}$$

so based on Lemma S2, the key is to derive $\varphi_N(V)$. From the chain rule, we have

$$\begin{aligned}
\dot{N}_\theta\Big|_{\theta=0} &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[\{p_{1,\theta}(X) - p_{0,\theta}(X)\}\mu_{11,\theta}(X)\right]\Big|_{\theta=0} \\
&= \mathbb{E}[\{p_1(X) - p_0(X)\}\mu_{11}(X)\mathrm{s}(X)] \quad \text{(Lemma S3)} \\
&\quad + \mathbb{E}_\theta\left[\{\dot{p}_{1,\theta}(X) - \dot{p}_{0,\theta}(X)\}\mu_{11,\theta}(X)\right]\big|_{\theta=0} \\
&\quad + \mathbb{E}_\theta\left[\{p_{1,\theta}(X) - p_{0,\theta}(X)\}\dot{\mu}_{11,\theta}(X)\right]\big|_{\theta=0}.
\end{aligned} \tag{S23}$$

Because $\mathbb{E}\{N\mathrm{s}(X)\} = 0$, the first term in (S23) equals

$$\mathbb{E}\{\{p_1(X) - p_0(X)\}\mu_{11}(X)\mathrm{s}(X)\} = \mathbb{E}\left(\left[\{p_1(X) - p_0(X)\}\mu_{11}(X) - N\right]\mathrm{s}(X)\}\right).$$

From Lemma S4, the second term in (S23) reduces to

$$\begin{aligned}
& \mathbb{E}_\theta\left[\{\dot{p}_{1,\theta}(X)-\dot{p}_{0,\theta}(X)\}\mu_{11,\theta}(X)\right]\big|_{\theta=0} \\
= \;& \mathbb{E}\left(\mathbb{E}\left[\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\mathrm{s}(S\mid Z,X)\mu_{11}(X)\mid X\right]\right) \\
= \;& \mathbb{E}\left[\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\mu_{11}(X)\mathrm{s}(S\mid Z,X)\right].
\end{aligned}$$

From Lemma S6, the third term in (S23) reduces to

$$\begin{aligned}
& \mathbb{E}_\theta\left[\{p_{1,\theta}(X)-p_{0,\theta}(X)\}\dot{\mu}_{11,\theta}(X)\right]\big|_{\theta=0} \\
= \;& \mathbb{E}\left[\{p_1(X)-p_0(X)\}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y\mid Z,S,X)\right].
\end{aligned}$$

Plugging the above three formulas into (S23) gives

$$\begin{aligned}
\dot{N}_\theta\Big|_{\theta=0} = \;& \mathbb{E}\left(\left[\{p_1(X)-p_0(X)\}\mu_{11}(X)-N\right]\mathrm{s}(X)\}\right) \\
& +\mathbb{E}\left[\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\mu_{11}(X)\mathrm{s}(S\mid Z,X)\right] \\
& +\mathbb{E}\left[\{p_1(X)-p_0(X)\}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y\mid Z,S,X)\right]. \qquad \text{(S24)}
\end{aligned}$$

We can verify that

$$\{p_1(X)-p_0(X)\}\mu_{11}(X)-N \in H_1, \qquad \text{(S25)}$$

$$\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\mu_{11}(X) \in H_3, \qquad \text{(S26)}$$

$$\{p_1(X)-p_0(X)\}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)} \in H_4. \qquad \text{(S27)}$$

Because H_1, H_2, H_3, and H_4 are orthogonal to each other, we can write (S24) as

$$\begin{aligned}
\dot{N}_\theta\Big|_{\theta=0} = \;& \mathbb{E}\left(\left[\{p_1(X)-p_0(X)\}\mu_{11}(X)-N\right]\mathrm{s}(V)\}\right) \\
& +\mathbb{E}\left[\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\mu_{11}(X)\mathrm{s}(V)\right] \\
& +\mathbb{E}\left[\{p_1(X)-p_0(X)\}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(V)\right].
\end{aligned}$$

As a result, we obtain the EIF for N:

$$\begin{aligned}
\varphi_N(V) = \;& \{p_1(X)-p_0(X)\}\mu_{11}(X)-N+\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\mu_{11}(X) \\
& +\{p_1(X)-p_0(X)\}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)} \\
= \;& (\psi_{S_1}-\psi_{S_0})\mu_{11}(X)-N+\left\{1-\frac{p_0(X)}{p_1(X)}\right\}\{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}\} \\
= \;& \frac{e_{10}(X)}{p_1(X)}\psi_{Y_1S_1}-N-\mu_{11}(X)\left\{\psi_{S_0}-\frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\},
\end{aligned}$$

which equals $\phi_{1,10} - N$ (see Theorem 2 for the expression of $\phi_{1,10}$). From Lemma S2, the EIF for $\mu_{1,10}$ is

$$\varphi_{1,10}(V) = \frac{1}{p_1 - p_0}\phi_{1,10} - \frac{\mu_{1,10}(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0}.$$

Similarly, the EIF for $\mu_{0,10}$ is

$$\varphi_{0,10}(V) = \frac{1}{p_1 - p_0}\phi_{0,10} - \frac{\mu_{0,10}(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0}.$$

Therefore, the EIF for τ_{10} is

$$\frac{(\phi_{1,10} - \phi_{0,10}) - \tau_{10}(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0}.$$

□

S6.2 Proof of Theorems S2, S3, and S4

The proofs are almost identical to the cases without restrictions. We omit the details due to the length of supplementary material. The earliest ArXiv version of our paper contains the details.

S6.3 Proof of Theorem S6

We will follow a similar route as in Section S6.1. The tangent space is the same as in (S20). To avoid new symbols, we use the same notation as in Section S6.1. We need the following lemma.

Lemma S7 *For $\omega_{1,10}(X)$, we have*

$$\dot{\omega}_{1,10,\theta}(X)|_{\theta=0} = \mathbb{E}\left[\left\{\frac{\omega_{1,10}(X)\psi_{S_1}}{p_1(X)} - \frac{\omega_{1,10}^2(X)(\psi_{S_1} - \psi_{S_0})}{p_1(X)} - \frac{\omega_{1,10}^2(X)\psi_{S_0}}{\epsilon_1(X)p_1(X)}\right\}\mathrm{s}(S \mid Z, X) \mid X\right].$$

Proof of Lemma S7. The parameter $\omega_{1,10}(X)$ has a ratio form

$$\omega_{1,10}(X) = \frac{\epsilon_1(X)\{e_{10}(X) + e_{11}(X)\}}{\epsilon_1(X)e_{10}(X) + e_{11}(X)} = \frac{\epsilon_1(X)p_1(X)}{\epsilon_1(X)\{p_1(X) - p_0(X)\} + p_0(X)}.$$

Thus, we can use Lemma S2 to calculate its path derivative with all distributions conditional on X. From Lemma S4, the numerator satisfies

$$\epsilon_1(X)\dot{p}_1(X)|_{\theta=0} = \mathbb{E}[\epsilon_1(X)\{\psi_{S_1} - p_1(X)\}\mathrm{s}(S \mid Z, X) \mid X],$$

and the denominator satisfies

$$\epsilon_1(X)\{\dot{p}_{1,\theta}(X) - \dot{p}_{0,\theta}(X)\}|_{\theta=0} + \dot{p}_{0,\theta}(X)|_{\theta=0} = \mathbb{E}\left[\{\epsilon_1(X)(\psi_{S_1} - \psi_{S_0} - e_{10}(X)) + \psi_{S_0} - p_0(X)\}\mathrm{s}(S \mid Z, X) \mid X\right].$$

Using Lemma S2, we can obtain $\dot{\omega}_{1,10,\theta}(X)|_{\theta=0} = \mathbb{E}\{\phi_{\omega,1,10}(V)\mathrm{s}(S \mid Z, X) \mid X\}$, where

$$\phi_{\omega,1,10}(V) \quad = \quad \frac{\epsilon_1(X)\{\psi_{S_1} - p_1(X)\} - \omega_{1,10}(X)\{\epsilon_1(X)(\psi_{S_1} - \psi_{S_0} - e_{10}(X)) + \psi_{S_0} - p_0(X)\}}{\epsilon_1(X)e_{10}(X) + e_{11}(X)}.$$

The result then follows from simple algebra.

□

We will derive the EIF for $\tau_{10} = \mu_{1,10} - \mu_{0,10}$ and omit the similar proofs of the EIFs for τ_{11} and τ_{00}. We can write $\mu_{1,10} = N/D$, where

$$N = \mathbb{E}\left[\{p_1(X) - p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)\right], \quad D = p_1 - p_0.$$

Lemma S5 implies that $\varphi_D(V) = (\psi_{S_1} - \psi_{S_0}) - D$ is the EIF for D, so based on Lemma S2, the key is to derive the EIF for N. From the chain rule, we have

$$\begin{aligned}
\dot{N}_\theta\Big|_{\theta=0} &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[\{p_{1,\theta}(X) - p_{0,\theta}(X)\}\omega_{1,10}(X)\mu_{11,\theta}(X)\right]\Big|_{\theta=0} \\
&= \mathbb{E}[\{p_1(X) - p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)\mathrm{s}(X)] \quad \text{(Lemma S3)} \\
&\quad + \mathbb{E}_\theta\left[\{\dot{p}_{1,\theta}(X) - \dot{p}_{0,\theta}(X)\}\omega_{1,10,\theta}(X)\mu_{11,\theta}(X)\right]|_{\theta=0} \\
&\quad + \mathbb{E}_\theta\left[\{p_{1,\theta}(X) - p_{0,\theta}(X)\}\dot{\omega}_{1,10,\theta}(X)\mu_{11,\theta}(X)\right]|_{\theta=0} \\
&\quad + \mathbb{E}_\theta\left[\{p_{1,\theta}(X) - p_{0,\theta}(X)\}\omega_{1,10,\theta}(X)\dot{\mu}_{11,\theta}(X)\right]|_{\theta=0}.
\end{aligned} \tag{S28}$$

Because $\mathbb{E}\{N\mathrm{s}(X)\} = 0$, the first term in (S28) equals

$$\begin{aligned}
&\mathbb{E}\{\{p_1(X) - p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)\mathrm{s}(X)\} \\
= \; &\mathbb{E}\left(\left[\{p_1(X) - p_0(X)\}\omega_{1,10}(X)\mu_{11}(X) - N\right]\mathrm{s}(X)\}\right).
\end{aligned}$$

From Lemma S4, the second term in (S28) reduces to

$$\begin{aligned}
&\mathbb{E}_\theta\left[\{\dot{p}_{1,\theta}(X) - \dot{p}_{0,\theta}(X)\}\omega_{1,10,\theta}(X)\mu_{11,\theta}(X)\right]|_{\theta=0} \\
= \; &\mathbb{E}\left(\mathbb{E}\left[\{\psi_{S_1} - \psi_{S_0} - p_1(X) + p_0(X)\}\mathrm{s}(S \mid Z, X)\omega_{1,10}(X)\mu_{11}(X) \mid X\right]\right) \\
= \; &\mathbb{E}\left[\{\psi_{S_1} - \psi_{S_0} - p_1(X) + p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)\mathrm{s}(S \mid Z, X)\right].
\end{aligned}$$

From Lemma S7, the third term in (S28) reduces to

$$\begin{aligned}
&\mathbb{E}_\theta\left[\{p_{1,\theta}(X) - p_{0,\theta}(X)\}\dot{\omega}_{1,10,\theta}(X)\mu_{11,\theta}(X)\right]|_{\theta=0} \\
= \; &\mathbb{E}\left[\{p_1(X) - p_0(X)\}\mu_{11}(X)\right. \\
&\quad \left. \cdot \left\{\frac{\omega_{1,10}(X)\psi_{S_1}}{p_1(X)} - \frac{\omega_{1,10}^2(X)(\psi_{S_1} - \psi_{S_0})}{p_1(X)} - \frac{\omega_{1,10}^2(X)\psi_{S_0}}{\epsilon_1(X)p_1(X)}\right\}\mathrm{s}(S \mid Z, X)\right]
\end{aligned}$$

From Lemma S6, the fourth term in (S28) reduces to

$$\begin{aligned}
&\mathbb{E}_\theta\left[\{p_{1,\theta}(X)-p_{0,\theta}(X)\}\omega_{1,10,\theta}(X)\dot{\mu}_{11,\theta}(X)\right]\big|_{\theta=0}\\
=\ &\mathbb{E}\left[\{p_1(X)-p_0(X)\}\omega_{1,10}(X)\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y\mid Z,S,X)\right].
\end{aligned}$$

Plugging the above four formulas into (S28) yields

$$\begin{aligned}
\dot{N}_\theta\Big|_{\theta=0} =\ & \mathbb{E}\left([\{p_1(X)-p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)-N]\,\mathrm{s}(X)\}\right)\\
&+\mathbb{E}\left[\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)\mathrm{s}(S\mid Z,X)\right]\\
&+\mathbb{E}\Bigg[\{p_1(X)-p_0(X)\}\mu_{11}(X)\\
&\quad\cdot\left\{\frac{\omega_{1,10}(X)\psi_{S_1}}{p_1(X)}-\frac{\omega_{1,10}^2(X)(\psi_{S_1}-\psi_{S_0})}{p_1(X)}-\frac{\omega_{1,10}^2(X)\psi_{S_0}}{\epsilon_1(X)p_1(X)}\right\}\mathrm{s}(S\mid Z,X)\Bigg]\\
&+\mathbb{E}\left[\{p_1(X)-p_0(X)\}\omega_{1,10}(X)\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y\mid Z,S,X)\right]. \qquad \text{(S29)}
\end{aligned}$$

We can verify that

$$\begin{aligned}
\{p_1(X)-p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)-N &\in H_1,\\
\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\omega_{1,10}(X)\mu_{11}(X) &\in H_3,\\
\{p_1(X)-p_0(X)\}\mu_{11}(X)\left\{\frac{\omega_{1,10}(X)\psi_{S_1}}{p_1(X)}-\frac{\omega_{1,10}^2(X)(\psi_{S_1}-\psi_{S_0})}{p_1(X)}-\frac{\omega_{1,10}^2(X)\psi_{S_0}}{\epsilon_1(X)p_1(X)}\right\} &\in H_3,\\
\{p_1(X)-p_0(X)\}\omega_{1,10}(X)\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)} &\in H_4.
\end{aligned}$$

Therefore, we can replace all the score functions in (S29) with $\mathrm{s}(V)$. As a result, the EIF for N is

$$\begin{aligned}
\varphi_N(V) =\ & \{p_1(X)-p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)-N+\{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)\}\omega_{1,10}(X)\mu_{11}(X)\\
&+\{p_1(X)-p_0(X)\}\mu_{11}(X)\left\{\frac{\omega_{1,10}(X)\psi_{S_1}}{p_1(X)}-\frac{\omega_{1,10}^2(X)(\psi_{S_1}-\psi_{S_0})}{p_1(X)}-\frac{\omega_{1,10}^2(X)\psi_{S_0}}{\epsilon_1(X)p_1(X)}\right\}\\
&+\{p_1(X)-p_0(X)\}\omega_{1,10}(X)\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\\
=\ & \frac{e_{10}(S)\omega_{1,10}(X)}{p_1(X)}\psi_{Y_1S_1}-N+\{\psi_{S_1}-\psi_{S_0}\}\omega_{1,10}(X)\mu_{11}(X)\\
&-e_{10}(X)\mu_{11}(X)\left\{\frac{\omega_{1,10}^2(X)(\psi_{S_1}-\psi_{S_0})}{p_1(X)}+\frac{\omega_{1,10}^2(X)\psi_{S_0}}{\epsilon_1(X)p_1(X)}\right\}\\
=\ & \frac{e_{10}(S)\omega_{1,10}(X)}{p_1(X)}\psi_{Y_1S_1}-N+\psi_{S_1}\mu_{11}(X)\left\{\omega_{1,10}(X)-\frac{e_{10}(X)\omega_{1,10}(X)}{p_1(X)}\right\}\\
&-\psi_{S_0}\mu_{11}(X)\left\{\omega_{1,10}(X)+\frac{e_{10}(X)\omega_{1,10}^2(X)}{\epsilon_1(X)p_1(X)}-\frac{e_{10}(X)\omega_{1,10}^2(X)}{p_1(X)}\right\}\\
=\ & \frac{e_{10}(S)\omega_{1,10}(X)}{p_1(X)}\psi_{Y_1S_1}-N-\frac{\omega_{1,10}^2(X)\mu_{11}(X)}{\epsilon_1(X)}\left\{\psi_{S_0}-\frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\},
\end{aligned}$$

which equals $\phi'_{1,10} - N$ (see Theorem S6 for the expression of $\phi'_{1,10}$). From Lemma S2, the EIF for $\mu_{1,10}$ is

$$\varphi_{1,10}(V) = \frac{\phi'_{1,10}}{p_1 - p_0} - \frac{\mu_{1,10}(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0}.$$

Similarly, we can obtain the EIF for $\mu_{0,10}$,

$$\varphi_{0,10}(V) = \frac{\phi'_{0,10}}{p_1 - p_0} - \frac{\mu_{0,10}(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0}.$$

Therefore, the EIF for τ_{10} is

$$\frac{(\phi'_{1,10} - \phi'_{0,10}) - \tau_{10}(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0}.$$

□

Solving $\mathbb{E}(\phi'_u) = 0$ for τ_u yields the following identification formulas for sensitivity analysis:

$$\tau_{10} = \frac{\mathbb{E}(\phi'_{1,10} - \phi'_{0,10})}{\mathbb{E}(\psi_{S_1} - \psi_{S_0})}, \quad \tau_{00} = \frac{\mathbb{E}(\phi'_{1,00} - \phi'_{0,00})}{\mathbb{E}(\psi_{1-S_1})}, \quad \tau_{11} = \frac{\mathbb{E}(\phi'_{1,11} - \phi'_{0,11})}{\mathbb{E}(\psi_{S_0})}.$$

Similar to the main text, we can construct the estimators for τ_u without principal ignorability.

S6.4 Proof of Theorem S9

We will derive the EIF for $\tau_{10} = \mu_{1,10} - \mu_{0,10}$ and omit the similar proofs of the EIFs for τ_{00}, τ_{11}, and τ_{01}. We can write $\mu_{1,10} = N/D$, where

$$N = \mathbb{E}\left\{\frac{p_1(X) - p_0(X)}{1 - \xi(X)}\mu_{11}(X)\right\}, \quad D = \mathbb{E}\left\{\frac{p_1(X) - p_0(X)}{1 - \xi(X)}\right\}.$$

Lemma S5 implies that

$$\phi_D(V) = \frac{\psi_{S_1} - \psi_{S_0}}{1 - \xi(X)} - D$$

is the EIF for D, so based on Lemma S2, the key is to derive the EIF for N. From the chain rule, we have

$$\begin{aligned} \dot{N}_\theta\Big|_{\theta=0} &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left\{\frac{p_{1,\theta}(X) - p_{0,\theta}(X)}{1 - \xi(X)}\mu_{11,\theta}(X)\right\}\Big|_{\theta=0} \\ &= \mathbb{E}\left\{\frac{p_1(X) - p_0(X)}{1 - \xi(X)}\mu_{11}(X)\mathrm{s}(X)\right\} \quad \text{(Lemma S3)} \\ &\quad + \mathbb{E}_\theta\left\{\frac{\dot{p}_{1,\theta}(X) - \dot{p}_{0,\theta}(X)}{1 - \xi(X)}\mu_{11,\theta}(X)\right\}\Big|_{\theta=0} \\ &\quad + \mathbb{E}_\theta\left\{\frac{p_{1,\theta}(X) - p_{0,\theta}(X)}{1 - \xi(X)}\dot{\mu}_{11,\theta}(X)\right\}\Big|_{\theta=0}. \end{aligned} \tag{S30}$$

Because $\mathbb{E}\{N\mathrm{s}(X)\}=0$, the first term in (S30) equals

$$\mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\mu_{11}(X)\mathrm{s}(X)\right\} \;=\; \mathbb{E}\left[\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\mu_{11}(X)-N\right\}\mathrm{s}(X)\right].$$

From Lemma S4, the second term in (S30) reduces to

$$\begin{aligned}
&\mathbb{E}_\theta\left\{\frac{\dot p_{1,\theta}(X)-\dot p_{0,\theta}(X)}{1-\xi(X)}\mu_{11,\theta}(X)\right\}\Big|_{\theta=0}\\
=&\;\mathbb{E}\left[\mathbb{E}\left\{\frac{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)}{1-\xi(X)}\mathrm{s}(S\mid Z,X)\mu_{11}(X)\mid X\right\}\right]\\
=&\;\mathbb{E}\left\{\frac{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)}{1-\xi(X)}\mu_{11}(X)\mathrm{s}(S\mid Z,X)\right\}.
\end{aligned}$$

From Lemma S6, the third term in (S30) reduces to

$$\begin{aligned}
&\mathbb{E}_\theta\left\{\frac{p_{1,\theta}(X)-p_{0,\theta}(X)}{1-\xi(X)}\dot\mu_{11,\theta}(X)\right\}\Big|_{\theta=0}\\
=&\;\mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y\mid Z,S,X)\right\}.
\end{aligned}$$

Plugging the above three formulas into (S30) gives

$$\begin{aligned}
\dot N_\theta\Big|_{\theta=0} =&\;\mathbb{E}\left[\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\mu_{11}(X)-N\right\}\mathrm{s}(X)\right]\\
&+\mathbb{E}\left\{\frac{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)}{1-\xi(X)}\mu_{11}(X)\mathrm{s}(S\mid Z,X)\right\}\\
&+\mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(Y\mid Z,S,X)\right\}.
\end{aligned}\tag{S31}$$

We can verify that

$$\begin{aligned}
\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\mu_{11}(X)-N\right\} &\in H_1,\\
\frac{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)}{1-\xi(X)}\mu_{11}(X) &\in H_3,\\
\frac{p_1(X)-p_0(X)}{1-\xi(X)}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)} &\in H_4.
\end{aligned}$$

Because H_1, H_2, H_3, and H_4 are orthogonal to each other, we can write (S31) as

$$\begin{aligned}
\dot N_\theta\Big|_{\theta=0} =&\;\mathbb{E}\left[\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\mu_{11}(X)-N\right\}\mathrm{s}(V)\right]\\
&+\mathbb{E}\left\{\frac{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)}{1-\xi(X)}\mu_{11}(X)\mathrm{s}(V)\right\}\\
&+\mathbb{E}\left\{\frac{p_1(X)-p_0(X)}{1-\xi(X)}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)}\mathrm{s}(V)\right\}.
\end{aligned}$$

As a result, we obtain the EIF for N:

$$\varphi_N(V) \;=\; \frac{p_1(X)-p_0(X)}{1-\xi(X)}\mu_{11}(X)-N+\frac{\psi_{S_1}-\psi_{S_0}-p_1(X)+p_0(X)}{1-\xi(X)}\mu_{11}(X)$$

$$
\begin{aligned}
& +\frac{p_1(X)-p_0(X)}{1-\xi(X)}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{p_1(X)} \\
= \;& \frac{\psi_{S_1}-\psi_{S_0}}{1-\xi(X)}\mu_{11}(X)-N+\left\{1-\frac{p_0(X)}{p_1(X)}\right\}\frac{\psi_{Y_1S_1}-\mu_{11}(X)\psi_{S_1}}{1-\xi(X)} \\
= \;& \frac{p_1(X)-p_0(X)}{\{1-\xi(X)\}p_1(X)}\psi_{Y_1S_1}-N-\frac{\mu_{11}(X)}{1-\xi(X)}\left\{\psi_{S_0}-\frac{p_0(X)}{p_1(X)}\psi_{S_1}\right\},
\end{aligned}
$$

which equals $\phi^*_{1,10}-N$ (see Theorem S9 for the expression of $\phi^*_{1,10}$). From Lemma S2, the EIF for $\mu_{1,10}$ is

$$
\varphi_{1,10}(V)=\frac{1}{e_{\xi,10}}\left\{\phi^*_{1,10}-\frac{\mu_{1,10}(\psi_{S_1}-\psi_{S_0})}{1-\xi(X)}\right\}.
$$

Similarly, the EIF for $\mu_{0,10}$ is

$$
\varphi_{0,10}(V)=\frac{1}{e_{\xi,10}}\left\{\phi^*_{0,10}-\frac{\mu_{0,10}(\psi_{S_1}-\psi_{S_0})}{1-\xi(X)}\right\}.
$$

Therefore, the EIF for τ_{10} is

$$
\frac{1}{e_{\xi,10}}\left\{(\phi^*_{1,10}-\phi^*_{0,10})-\frac{\tau_{10}(\psi_{S_1}-\psi_{S_0})}{1-\xi(X)}\right\}.
$$

□

S7 Proofs of the multiple robustness and local efficiency

In this section, we prove Theorems 3, 4, S7, and S10.

S7.1 Proof of Theorem 3

We prove the triple robustness and semiparametric efficiency for $\widehat{\tau}_{10}$. The proofs for $\widehat{\tau}_{00}$ and $\widehat{\tau}_{11}$ are similar and hence omitted.

Proof of the triple robustness. As discussed in the main text, the proof of the triple robustness reduces to the calculation of the probability limit of $\mathbb{P}_n(\widehat{\phi}_{1,10})-\mathbb{E}\{Y_1\mathbf{1}(U=10)\}$. From the equivalence relationship in (S4), $\mathbb{P}_n(\widehat{\phi}_{1,10})$ can be rewritten as

$$
\mathbb{P}_n\left[e_{10}(X;\widehat{\gamma})\frac{S}{p_1(X;\widehat{\gamma})}\frac{Z}{\pi(X;\widehat{\alpha})}\{Y-\mu_{11}(X;\widehat{\beta})\}\right]+\mathbb{P}_n\left\{\widehat{\psi}_{\mu_{11}(X;\widehat{\beta})S_1}\right\}-\mathbb{P}_n\left\{\widehat{\psi}_{\mu_{11}(X;\widehat{\beta})S_0}\right\}. \quad \text{(S32)}
$$

The first term in (S32) is consistent for

$$
\begin{aligned}
& \mathbb{E}\left[e_{10}(X;\gamma^*)\frac{S}{p_1(X;\gamma^*)}\frac{Z}{\pi(X;\alpha^*)}\{Y-\mu_{11}(X;\beta^*)\}\right] \\
= \;& \mathbb{E}\left(\mathbb{E}\left[e_{10}(X;\gamma^*)\frac{\mathbb{P}(Z=1,S=1\mid X)}{p_1(X;\gamma^*)\pi(X;\alpha^*)}\{Y-\mu_{11}(X;\beta^*)\}\mid Z=1,S=1,X\right]\right) \quad \text{(LOTE)} \\
= \;& \mathbb{E}\left[\frac{\{p_1(X;\gamma^*)-p_0(X;\gamma^*)\}p_1(X)\pi(X)}{p_1(X;\gamma^*)\pi(X;\alpha^*)}\{\mu_{11}(X)-\mu_{11}(X;\beta^*)\}\right];
\end{aligned}
$$

the second term in (S32) is consistent for

$$\begin{aligned}
&\mathbb{E}\left[\frac{\{\mu_{11}(X;\beta^*)S_1-\mu_{11}(X;\beta^*)p_1(X;\gamma^*)\}\mathbf{1}(Z=1)}{\pi(X;\alpha^*)}+\mu_{11}(X;\beta^*)p_1(X;\gamma^*)\right]\\
=\ &\mathbb{E}\left[\frac{\{\mu_{11}(X;\beta^*)p_1(X)-\mu_{11}(X;\beta^*)p_1(X;\gamma^*)\}\pi(X)}{\pi(X;\alpha^*)}+\mu_{11}(X;\beta^*)p_1(X;\gamma^*)\right]; \quad \text{(LOTE)}
\end{aligned}$$

similarly, the third term in (S32) is consistent for

$$\begin{aligned}
&\mathbb{E}\left[\frac{\{\mu_{11}(X;\beta^*)S_0-\mu_{11}(X;\beta^*)p_0(X;\gamma^*)\}\mathbf{1}(Z=0)}{1-\pi(X;\alpha^*)}+\mu_{11}(X;\beta^*)p_0(X;\gamma^*)\right]\\
=\ &\mathbb{E}\left[\frac{\{\mu_{11}(X;\beta^*)p_0(X)-\mu_{11}(X;\beta^*)p_0(X;\gamma^*)\}\{1-\pi(X)\}}{1-\pi(X;\alpha^*)}+\mu_{11}(X;\beta^*)p_0(X;\gamma^*)\right]. \quad \text{(LOTE)}
\end{aligned}$$

For the ease of disposition, we suppress the dependence of the functions on X and write $(\pi(X),\pi(X;\alpha^*))$ as (π,π^*), $(p_z(X),p_z(X;\gamma^*))$ as (p_z,p_z^*), and $(\mu_{11}(X),\mu_{11}(X;\beta^*))$ as (μ_{11},μ_{11}^*). Therefore, $\mathbb{P}_n(\widehat{\phi}_{1,10})-\mathbb{E}\{Y_1\mathbf{1}(U=10)\}$ is consistent for

$$\begin{aligned}
&\mathbb{E}\left\{\frac{\pi(p_1^*-p_0^*)p_1(\mu_{11}-\mu_{11}^*)}{\pi^*p_1^*}\right\}+\mathbb{E}\left\{\frac{(\mu_{11}^*p_1-\mu_{11}^*p_1^*)\pi}{\pi^*}+\mu_{11}^*p_1^*\right\}\\
&-\mathbb{E}\left\{\frac{(\mu_{11}^*p_0-\mu_{11}^*p_0^*)(1-\pi)}{1-\pi^*}+\mu_{11}^*p_0^*\right\}-\mu_{11}(p_1-p_0)\\
=\ &\mathbb{E}\left\{\frac{\pi p_1(\mu_{11}-\mu_{11}^*)}{\pi^*}+\frac{(\mu_{11}^*p_1-\mu_{11}^*p_1^*)\pi}{\pi^*}+\mu_{11}^*p_1^*-\mu_{11}p_1\right\}\\
&-\mathbb{E}\left\{\frac{\pi p_0^*p_1(\mu_{11}-\mu_{11}^*)}{\pi^*p_1^*}+\frac{(\mu_{11}^*p_0-\mu_{11}^*p_0^*)(1-\pi)}{1-\pi^*}+\mu_{11}^*p_0^*-\mu_{11}p_0\right\}. \quad \text{(S33)}
\end{aligned}$$

We simplify the terms in the two expectations of (S33) separately. The first term reduces to

$$\begin{aligned}
\frac{\pi p_1(\mu_{11}-\mu_{11}^*)}{\pi^*}+\frac{(\mu_{11}^*p_1-\mu_{11}^*p_1^*)\pi}{\pi^*}+\mu_{11}^*p_1^*-\mu_{11}p_1 &= \frac{\pi p_1\mu_{11}}{\pi^*}-\frac{\mu_{11}^*p_1^*\pi}{\pi^*}+\mu_{11}^*p_1^*-\mu_{11}p_1\\
&= \frac{(\mu_{11}p_1-\mu_{11}^*p_1^*)(\pi-\pi^*)}{\pi^*}.
\end{aligned}$$

The second term reduces to

$$\begin{aligned}
&\frac{\pi p_0^*p_1(\mu_{11}-\mu_{11}^*)}{\pi^*p_1^*}+\frac{(\mu_{11}^*p_0-\mu_{11}^*p_0^*)(1-\pi)}{1-\pi^*}+\mu_{11}^*p_0^*-\mu_{11}p_0\\
=\ &\frac{(1-\pi^*)\pi p_0^*p_1(\mu_{11}-\mu_{11}^*)+(1-\pi)\pi^*p_1^*(\mu_{11}^*p_0-\mu_{11}^*p_0^*)+\pi^*p_1^*(1-\pi^*)(\mu_{11}^*p_0^*-\mu_{11}p_0)}{\pi^*p_1^*(1-\pi^*)}\\
=\ &\frac{\pi p_0^*p_1(\mu_{11}-\mu_{11}^*)+\pi^*p_1^*(\mu_{11}^*p_0-\mu_{11}^*p_0^*)+\pi^*p_1^*(\mu_{11}^*p_0^*-\mu_{11}p_0)}{\pi^*p_1^*(1-\pi^*)}\\
&-\frac{\pi p_0^*p_1(\mu_{11}-\mu_{11}^*)+\pi p_1^*(\mu_{11}^*p_0-\mu_{11}^*p_0^*)+\pi^*p_1^*(\mu_{11}^*p_0^*-\mu_{11}p_0)}{p_1^*(1-\pi^*)}\\
=\ &\frac{\pi p_0^*p_1(\mu_{11}-\mu_{11}^*)+\pi^*p_1^*(\mu_{11}^*p_0-\mu_{11}p_0)}{\pi^*p_1^*(1-\pi^*)}\\
&-\frac{(\pi p_0^*p_1-\pi^*p_0p_1^*)\mu_{11}+(\pi p_0p_1^*-\pi p_0^*p_1^*-\pi p_0^*p_1+\pi^*p_0^*p_1^*)\mu_{11}^*}{p_1^*(1-\pi^*)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(\pi p_0^* p_1 - \pi^* p_0 p_1^*)(\mu_{11} - \mu_{11}^*)}{\pi^* p_1^*(1-\pi^*)} - \frac{(\pi p_0^* p_1 - \pi^* p_0 p_1^*)\mu_{11} - (\pi p_0^* p_1 - \pi^* p_0 p_1^*)\mu_{11}^*}{p_1^*(1-\pi^*)} \\
&\quad - \frac{(\pi p_0 p_1^* - \pi p_0^* p_1^* - \pi^* p_0 p_1^* + \pi^* p_0^* p_1^*)\mu_{11}^*}{p_1^*(1-\pi^*)} \\
&= \frac{(\pi p_0^* p_1 - \pi^* p_0 p_1^*)(\mu_{11} - \mu_{11}^*)}{\pi^* p_1^*} - \frac{(\pi - \pi^*)(p_0 - p_0^*)\mu_{11}^*}{1-\pi^*}.
\end{aligned}
$$

Combining them yields the bias formulas in the main text.

Proof of the semiparametric efficiency. Under $\mathcal{M}_{\text{tp+ps+om}}$, we show that the influence function of $\widehat{\tau}_{10}$ is the same as the EIF in Theorem 2 and therefore $\widehat{\tau}_{10}$ achieves the local efficiency. Let θ denote the parameter vector containing α, β, and γ. Let θ^* be the probability limit of $\widehat{\theta}$. We first establish that for a ratio estimator $\mathbb{P}_n N(V;\widehat{\theta})/\mathbb{P}_n D(V;\widehat{\theta})$, by a Taylor expansion,

$$
\frac{\mathbb{P}_n N(V;\widehat{\theta})}{\mathbb{P}_n D(V;\widehat{\theta})} = \frac{\mathbb{P}_n N(V;\widehat{\theta})}{\mathbb{P}\{D(V;\theta^*)\}} - \frac{\mathbb{P}\{N(V;\theta^*)\}}{[\mathbb{P}\{D(V;\theta^*)\}]^2}\left[\mathbb{P}_n D(V;\widehat{\theta}) - \mathbb{P}\{D(V;\theta^*)\}\right] + o_{\mathbb{P}}(n^{-1/2}). \tag{S34}
$$

For $\widehat{\tau}_{10} = \mathbb{P}_n(\widehat{\phi}_{1,10} - \widehat{\phi}_{0,10})/\mathbb{P}_n(\widehat{\psi}_{S_1} - \widehat{\psi}_{S_0})$, we have $N(V;\theta) = \phi_{1,10}(\theta) - \phi_{0,10}(\theta)$ and $D(V;\theta) = \psi_{S_1}(\theta) - \psi_{S_0}(\theta)$, where

$$
\begin{aligned}
\phi_{1,10}(\theta) &= \frac{e_{10}(X;\gamma)}{p_1(X;\gamma)}\frac{Z}{\pi(X;\alpha)}\{Y - \mu_{11}(X;\beta)\}S + \psi_{\mu_{11}(X;\beta)S_1} - \psi_{\mu_{11}(X;\beta)S_0}, \\
\phi_{0,10}(\theta) &= \frac{e_{10}(X;\gamma)}{1-p_0(X;\gamma)}\frac{1-Z}{1-\pi(X;\alpha)}\{Y - \mu_{00}(X;\beta)\}(1-S) + \psi_{\mu_{00}(X;\beta)(1-S_0)} - \psi_{\mu_{00}(X;\beta)(1-S_1)}, \\
\psi_{S_z}(\theta) &= \frac{\{S - p_z(X;\gamma)\}\mathbf{1}(Z=z)}{\mathbb{P}(Z=z \mid X;\alpha)} + p_z(X;\gamma) \ (z=0,1).
\end{aligned}
$$

Under $\mathcal{M}_{\text{tp+ps+om}}$, $\phi_{1,10}(\theta^*) = \phi_{1,10}(V)$, $\phi_{0,10}(\theta^*) = \phi_{0,10}(V)$, $\psi_{S_z}(\theta^*) = \psi_{S_z}$ $(z=0,1)$, $\mathbb{P}\{N(V;\theta^*)\} = \tau_{10}(p_1 - p_0)$, and $\mathbb{P}\{D(V;\theta^*)\} = p_1 - p_0$. By Taylor expansions, we have

$$
\begin{aligned}
\mathbb{P}_n N(V;\widehat{\theta}) &= \mathbb{P}_n N(V;\theta^*) + \mathbb{P}\left\{\dot{N}(V;\theta^*)\right\}(\widehat{\theta} - \theta) + o_{\mathbb{P}}(n^{-1/2}), &\text{(S35)}\\
\mathbb{P}_n D(V;\widehat{\theta}) &= \mathbb{P}_n D(V;\theta^*) + \mathbb{P}\left\{\dot{D}(V;\theta^*)\right\}(\widehat{\theta} - \theta) + o_{\mathbb{P}}(n^{-1/2}). &\text{(S36)}
\end{aligned}
$$

By some algebra, we can verify that $\mathbb{P}\left\{\dot{N}(V;\theta^*)\right\} = \mathbb{P}\left\{\dot{D}(V;\theta^*)\right\} = 0$. Note that (S35) and (S36) hold under the sufficient conditions that (i) the propensity score and its estimator are bounded away from zero and one and (ii) the principal scores $p_1(X;\gamma^*)$ and $1 - p_0(X;\gamma^*)$ and their estimators are bounded away from zero almost surely. When condition (i) is violated, $\mathbb{P}_n N(V;\widehat{\theta})$ and $\mathbb{P}_n D(V;\widehat{\theta})$ can be unbounded. When condition (ii) is violated, $\mathbb{P}\left\{\dot{N}(V;\theta^*)\right\}$ and $\mathbb{P}\left\{\dot{D}(V;\theta^*)\right\}$ may be undefined. Combining (S34), (S35), and (S36), we obtain

$$
\hat{\tau}_{10} - \tau_{10} = \frac{\mathbb{P}_n N(V;\widehat{\theta})}{\mathbb{P}_n D(V;\widehat{\theta})} - \tau_{10}
$$

$$
\begin{aligned}
&= \frac{\mathbb{P}_n N(V;\widehat{\theta})}{\mathbb{P}\{D(V;\theta^*)\}} - \frac{\mathbb{P}\{N(V;\theta^*)\}}{[\mathbb{P}\{D(V;\theta^*)\}]^2}\left[\mathbb{P}_n D(V;\widehat{\theta}) - \mathbb{P}\{D(V;\theta^*)\}\right] - \tau_{10} + o_{\mathbb{P}}(n^{-1/2}) \\
&= \frac{\mathbb{P}_n N(V;\theta^*)}{\mathbb{P}\{D(V;\theta^*)\}} - \frac{\mathbb{P}\{N(V;\theta^*)\}}{[\mathbb{P}\{D(V;\theta^*)\}]^2}[\mathbb{P}_n D(V;\theta^*) - \mathbb{P}\{D(V;\theta^*)\}] - \tau_{10} + o_{\mathbb{P}}(n^{-1/2}) \\
&= \frac{\mathbb{P}_n N(V;\theta^*)}{\mathbb{P}\{D(V;\theta^*)\}} - \frac{\mathbb{P}\{N(V;\theta^*)\}}{[\mathbb{P}\{D(V;\theta^*)\}]^2}[\mathbb{P}_n D(V;\theta^*) - \mathbb{P}\{D(V;\theta^*)\}] - \tau_{10} + o_{\mathbb{P}}(n^{-1/2}) \\
&= \frac{\mathbb{P}_n N(V;\theta^*)}{\mathbb{P}\{D(V;\theta^*)\}} - \tau_{10}\frac{\mathbb{P}_n D(V;\theta^*)}{\mathbb{P}\{D(V;\theta^*)\}} + o_{\mathbb{P}}(n^{-1/2}) \\
&= \frac{\mathbb{P}_n(\phi_{1,10} - \phi_{0,10})}{p_1 - p_0} - \tau_{10}\frac{\mathbb{P}_n(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0} + o_{\mathbb{P}}(n^{-1/2}) \\
&= \mathbb{P}_n \frac{\phi_{1,10} - \phi_{0,10} - \tau_{10}(\psi_{S_1} - \psi_{S_0})}{p_1 - p_0} + o_{\mathbb{P}}(n^{-1/2}) \\
&= \mathbb{P}_n \phi_{10} + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

This completes the proof. □

S7.2 Proof of Theorem 4

We show that $\widehat{\tau}_{10}$ using $\{\widehat{\pi}(x), \widehat{p}_z(x), \widehat{\mu}_{zs}(x)\}$ satisfying the regularity conditions (a)–(d) in Theorem 4 is asymptotically normal and has the influence function ϕ_{10} as in Theorem 2, therefore achieving the semiparametric efficiency. The proofs for $\widehat{\tau}_{00}$ and $\widehat{\tau}_{11}$ are similar and hence omitted. Let θ denote the nuisance functions $\{\pi(x), p_z(x), \mu_{zs}(x)\}$, abbreviated to (π, p_z, μ_{zs}) for simplicity. Let θ^* be the probability limit of $\widehat{\theta}$.

For $\widehat{\tau}_{10} = \mathbb{P}_n(\widehat{\phi}_{1,10} - \widehat{\phi}_{0,10})/\mathbb{P}_n(\widehat{\psi}_{S_1} - \widehat{\psi}_{S_0})$, it is a ratio estimator $\mathbb{P}_n N(V;\widehat{\theta})/\mathbb{P}_n D(V;\widehat{\theta})$ where $N(V;\theta) = \phi_{1,10}(\theta) - \phi_{0,10}(\theta)$ and $D(V;\theta) = \psi_{S_1}(\theta) - \psi_{S_0}(\theta)$ with

$$
\begin{aligned}
\phi_{1,10}(\theta) &= \frac{e_{10}(X)}{p_1(X)}\frac{Z}{\pi(X)}\{Y - \mu_{11}(X)\}S + \psi_{\mu_{11}(X)S_1} - \psi_{\mu_{11}(X)S_0}, \\
\phi_{0,10}(\theta) &= \frac{e_{10}(X)}{1 - p_0(X)}\frac{1 - Z}{1 - \pi(X)}\{Y - \mu_{00}(X)\}(1 - S) + \psi_{\mu_{00}(X)(1-S_0)} - \psi_{\mu_{00}(X)(1-S_1)}, \\
\psi_{S_z}(\theta) &= \frac{\{S - p_z(X)\}\mathbf{1}(Z = z)}{\mathbb{P}(Z = z \mid X)} + p_z(X).
\end{aligned}
$$

We continue with the Taylor expansion of a ratio estimator as in (S34). Condition (a) implies that $\theta^* = \{\pi(x), p_z(x), \mu_{zs}(x)\}$ and thus $\phi_{1,10}(\theta^*) = \phi_{1,10}(V)$, $\phi_{0,10}(\theta^*) = \phi_{0,10}(V)$, and $\psi_{S_z}(\theta^*) = \psi_{S_z}$ $(z = 0, 1)$. By the empirical process theory, we have

$$
\begin{aligned}
\mathbb{P}_n N(V;\widehat{\theta}) - \mathbb{P}N(V;\theta^*) &= (\mathbb{P}_n - \mathbb{P})N(V;\widehat{\theta}) + \mathbb{P}\{N(V;\widehat{\theta}) - N(V;\theta^*)\} \\
&= (\mathbb{P}_n - \mathbb{P})N(V;\theta^*) + \mathbb{P}\{N(V;\widehat{\theta}) - N(V;\theta^*)\} + o_{\mathbb{P}}(n^{-1/2}), \quad \text{(S37)}
\end{aligned}
$$

where the second equality follows by Condition (b). It remains to analyze the second term $\mathbb{P}\{N(V;\widehat{\theta}) -$

$N(V;\theta^*)\}$. Following the derivation of the bias formula, we have

$$
\begin{aligned}
|\mathbb{P}\{N(V;\widehat{\theta})-N(V;\theta^*)\}| &= \Bigg|\mathbb{P}\left[\frac{\{\mu_{11}(X)p_1(X)-\widehat{\mu}_{11}(X)\widehat{p}_1(X)\}\{\pi(X)-\widehat{\pi}(X)\}}{\widehat{\pi}(X)}\right] \\
&\quad +\mathbb{P}\left[\frac{\{\pi(X)\widehat{p}_0(X)p_1(X)-\widehat{\pi}(X)p_0(X)\widehat{p}_1(X)\}\{\mu_{11}(X)-\widehat{\mu}_{11}(X)\}}{\widehat{\pi}(X)\widehat{p}_1(X)}\right] \\
&\quad +\mathbb{P}\left[\frac{\{\pi(X)-\widehat{\pi}(X)\}\{p_0(X)-\widehat{p}_0(X)\}\widehat{\mu}_{11}(X)}{1-\widehat{\pi}(X)}\right]\Bigg| \\
&\leq \left|\mathbb{P}\left[\frac{\{\mu_{11}(X)-\widehat{\mu}_{11}(X)\}p_1(X)\{\pi(X)-\widehat{\pi}(X)\}}{\widehat{\pi}(X)}\right]\right| \\
&\quad +\left|\mathbb{P}\left[\frac{\{p_1(X)-\widehat{p}_1(X)\}\widehat{\mu}_{11}(X)\{\pi(X)-\widehat{\pi}(X)\}}{\widehat{\pi}(X)}\right]\right| \\
&\quad +\left|\mathbb{P}\left[\frac{\{\widehat{p}_0(X)-p_0(X)\}\pi(X)p_1(X)\{\mu_{11}(X)-\widehat{\mu}_{11}(X)\}}{\widehat{\pi}(X)\widehat{p}_1(X)}\right]\right| \\
&\quad +\left|\mathbb{P}\left[\frac{\{\pi(X)-\widehat{\pi}(X)\}p_1(X)p_0(X)\{\mu_{11}(X)-\widehat{\mu}_{11}(X)\}}{\widehat{\pi}(X)\widehat{p}_1(X)}\right]\right| \\
&\quad +\left|\mathbb{P}\left[\frac{\{p_1(X)-\widehat{p}_1(X)\}\widehat{\pi}(X)p_0(X)\{\mu_{11}(X)-\widehat{\mu}_{11}(X)\}}{\widehat{\pi}(X)\widehat{p}_1(X)}\right]\right| \\
&\quad +\left|\mathbb{P}\left[\frac{\{\pi(X)-\widehat{\pi}(X)\}\{p_0(X)-\widehat{p}_0(X)\}\widehat{\mu}_{11}(X)}{1-\widehat{\pi}(X)}\right]\right|.
\end{aligned}
$$

By the Cauchy–Schwarz inequality and Conditions (a), (c) and (d), it follows that for some constant C, we have

$$
\begin{aligned}
|\mathbb{P}\{N(V;\widehat{\theta})-N(V;\theta^*)\}| &\leq C\times||\mu_{11}(X)-\widehat{\mu}_{11}(X)||_2\times\{||\pi(X)-\widehat{\pi}(X)||_2 \\
&\quad +||p_1(X)-\widehat{p}_1(X)||_2+||p_0(X)-\widehat{p}_0(X)||_2\} \\
&\quad +C\times||\pi(X)-\widehat{\pi}(X)||_2\times\{||p_1(X)-\widehat{p}_1(X)||_2+||p_0(X)-\widehat{p}_0(X)||_2\} \\
&= o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}
$$

Continuing with (S37), we have $\mathbb{P}_nN(V;\widehat{\theta})-\mathbb{P}N(V;\theta^*)=(\mathbb{P}_n-\mathbb{P})N(V;\theta^*)+o_{\mathbb{P}}(n^{-1/2})$. Similarly, under Conditions (a)–(d), $\mathbb{P}_nD(V;\widehat{\theta})-\mathbb{P}D(V;\theta^*)=(\mathbb{P}_n-\mathbb{P})D(V;\theta^*)+o_{\mathbb{P}}(n^{-1/2})$. Plugging these into (S34), we obtain $\hat{\tau}_{10}-\tau_{10}=\mathbb{P}_n\phi_{10}+o_{\mathbb{P}}(n^{-1/2})$. This completes the proof. □

S7.3 Proof of Theorem S7

We prove the double robustness for $\widehat{\tau}'_{10}$. The proofs for $\widehat{\tau}'_{00}$ and $\widehat{\tau}'_{11}$ are similar and hence omitted.

We write $\widehat{\tau}'_{10}=\mathbb{P}_n(\widehat{\phi}'_{1,10}-\widehat{\phi}'_{0,10})/\mathbb{P}_n(\widehat{\psi}_{S_1}-\widehat{\psi}_{S_0})$. The denominator is consistent for $\mathbb{E}(S_1-S_0)=\mathbb{P}(U=10)$ under $\mathcal{M}_{\text{ps}}$, which is a superset of $\mathcal{M}_{\text{tp+ps}}\cup\mathcal{M}_{\text{ps+om}}$. For the numerator, we will show that under $\mathcal{M}_{\text{ps}}$, $\mathbb{P}_n(\widehat{\phi}_{1,10})-\mathbb{E}\{Y_1\mathbf{1}(U=10)\}$ has the probability limit

$$
\mathbb{E}\left[\frac{\omega_{1,10}(X)e_{10}(X)\{\mu_{11}(X)-\mu_{11}(X;\beta^*)\}\{\pi(X)-\pi(X;\alpha^*)\}}{\pi(X;\alpha^*)}\right], \tag{S38}
$$

which equals 0 if $\mathcal{M}_{\text{tp}} \cup \mathcal{M}_{\text{om}}$ further holds. As a result, $\mathbb{P}_n(\widehat{\phi}'_{1,10})$ is consistent for $\mathbb{E}\{Y_1 \mathbf{1}(U=10)\}$ under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{ps+om}}$. Similarly, we can show $\mathbb{P}_n(\widehat{\phi}'_{0,10})$ is consistent for $\mathbb{E}\{Y_0 \mathbf{1}(U=10)\}$ under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{ps+om}}$, which leads to the double robustness of $\widehat{\tau}'_{10}$. Below we prove (S38).

We can write $\mathbb{P}_n(\widehat{\phi}'_{1,10})$ as

$$\mathbb{P}_n\left\{\frac{\omega_{1,10}(X;\widehat{\gamma})e_{10}(X;\widehat{\gamma})}{p_1(X;\widehat{\gamma})}\widehat{\psi}_{Y_1S_1}\right\} - \mathbb{P}_n\left[\frac{\omega_{1,10}^2(X;\widehat{\gamma})\mu_{11}(X;\widehat{\beta})}{\epsilon_1(X)}\left\{\widehat{\psi}_{S_0} - \frac{p_0(X;\widehat{\gamma})}{p_1(X;\widehat{\gamma})}\widehat{\psi}_{S_1}\right\}\right]. \tag{S39}$$

The first term of (S39) is consistent for

$$\begin{aligned}
& \mathbb{E}\left(\frac{\omega_{1,10}(X)e_{10}(X)}{p_1(X)}\left[\frac{\{\mu_{11}(X)p_1(X) - \mu_{11}(X;\beta^*)p_1(X)\}\pi(X)}{\pi(X;\alpha^*)} + \mu_{11}(X;\beta^*)p_1(X)\right]\right) \\
= \ & \mathbb{E}\left(\omega_{1,10}(X)e_{10}(X)\left[\frac{\{\mu_{11}(X) - \mu_{11}(X;\beta^*)\}\pi(X)}{\pi(X;\alpha^*)} + \mu_{11}(X;\beta^*)\right]\right).
\end{aligned}$$

The second term of (S39) is consistent for

$$\begin{aligned}
& \mathbb{E}\left(\frac{\omega_{1,10}^2(X)\mu_{11}(X;\beta^*)}{\epsilon_1(X)}\left[\frac{\{S_0 - p_0(X)\}\mathbf{1}(Z=0)}{1-\pi(X;\alpha^*)} + p_0(X)\right]\right) \\
& + - \mathbb{E}\left(\frac{\omega_{1,10}^2(X)\mu_{11}(X;\beta^*)p_0(X)}{\epsilon_1(X)p_1(X)}\left[\frac{\{S_1 - p_1(X)\}\mathbf{1}(Z=1)}{\pi(X;\alpha^*)} + p_1(X)\right]\right) \\
= \ & \mathbb{E}\left\{\frac{\omega_{1,10}^2(X)\mu_{11}(X;\beta^*)p_0(X)}{\epsilon_1(X)}\right\} - \mathbb{E}\left\{\frac{\omega_{1,10}^2(X)\mu_{11}(X;\beta^*)p_0(X)p_1(X)}{\epsilon_1(X)p_1(X)}\right\} \quad \text{(LOTE)} \\
= \ & 0.
\end{aligned}$$

These two terms, coupled with the fact that $\mathbb{E}\{Y_1\mathbf{1}(U=10)\} = \mathbb{E}\{\omega_{1,10}(X)\mu_{11}(X)e_{10}(X)\}$ from (S12), imply (S38).

The proof of the semiparametric efficiency is similar to that of Theorem 3 and hence omitted. □

S7.4 Proof of Theorem S10

We prove the double robustness for $\widehat{\tau}^*_{10}$. The proofs for the other three estimators are similar and hence omitted.

We write

$$\widehat{\tau}^*_{10} = \mathbb{P}_n(\widehat{\phi}^*_{1,10} - \widehat{\phi}^*_{0,10})/\mathbb{P}_n\left(\frac{\widehat{\psi}_{S_1} - \widehat{\psi}_{S_0}}{1-\xi(X)}\right).$$

We can show that the denominator has the probability limit

$$\mathbb{E}\left(\frac{[\{p_1(X;\gamma^*) - p_1(X)\} - \{p_0(X;\gamma^*) - p_0(X)\}]\{\pi(X;\alpha^*) - \pi(X)\}}{\{1-\xi(X)\}\pi(X;\alpha^*)}\right) + \mathbb{E}\left\{\frac{p_1(X) - p_0(X)}{1-\xi(X)}\right\}.$$

The first term is equal to zero under $\mathcal{M}_{\text{tp}} \cup \mathcal{M}_{\text{ps}}$ and the second term is equal to $\mathbb{P}(U=10)$ under (S6). Therefore, the denominator is consistent for $\mathbb{P}(U=10)$ under $\mathcal{M}_{\text{tp}} \cup \mathcal{M}_{\text{ps}}$, which is a superset of $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$.

From the derivation of the probability limit of the denominator, we see that the only difference between the derivation here and that in Theorem 3 is the additional term $1-\xi(X)$ in the expectation. Because $\xi(X)$ is known, this does not change the formulas in the proof of Theorem 3 other than adding $1-\xi(X)$ in the denominators. Therefore, we can show that $\mathbb{P}_n(\widehat{\phi}^*_{1,10}) - \mathbb{E}\{Y_1 \mathbf{1}(U=10)\}$ is consistent for

$$\mathbb{E}\left\{ \frac{(\mu_{11}p_1 - \mu^*_{11}p^*_1)(\pi - \pi^*)}{(1-\xi)\pi^*} - \frac{(\pi p^*_0 p_1 - \pi^* p_0 p^*_1)(\mu_{11} - \mu^*_{11})}{(1-\xi)\pi^* p^*_1} + \frac{(\pi - \pi^*)(p_0 - p^*_0)\mu^*_{11}}{(1-\xi)(1-\pi^*)} \right\},$$

where we suppress the dependence of the functions on X for the ease of disposition. As a result, $\mathbb{P}_n(\widehat{\phi}^*_{1,10})$ is consistent for $\mathbb{E}\{Y_1\mathbf{1}(U=10)\}$ under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$. Similarly, we can show $\mathbb{P}_n(\widehat{\phi}^*_{0,10})$ is consistent for $\mathbb{E}\{Y_0\mathbf{1}(U=10)\}$ under $\mathcal{M}_{\text{tp+ps}} \cup \mathcal{M}_{\text{tp+om}} \cup \mathcal{M}_{\text{ps+om}}$, which leads to the triple robustness of $\widehat{\tau}^*_{10}$.

The proof of the semiparametric efficiency is similar to that of Theorem 3 and hence omitted. □

References

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*, 962–973.

Bickel, P. J., C. Klaassen, Y. Ritov, and J. Wellner (1993). *Efficient and Adaptive Inference in Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Ding, P. and J. Lu (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*, 757–777.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica 79*, 437–452.

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

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*, 243–263.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*, 523–539.

Little, R. and H. An (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 949–968.

Lumley, T., P. A. Shaw, and J. Y. Dai (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review 79*, 200–220.

Robins, J. M. and A. Rotnitzky (1998). Discussion of "Robust models in probability sampling" by Firth and Bennett. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*, 51–52.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*, 1096–1120.

Shen, C., X. Li, and L. Li (2014). Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in Medicine 33*, 555–568.

Tan, Z. (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika 107*, 137–158.

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

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*, 910–922.