Biometrics 000, 000–000 DOI: 000
March 2022

Supporting Information for "Instrumented Difference-in-Differences"

Ting Ye[1,*], Ashkan Ertefaie[2], James Flory[3], Sean Hennessy[4] and Dylan S. Small[5]

[1]Department of Biostatistics, University of Washington.

[2]Department of Biostatistics and Computational Biology, University of Rochester.

[3]Department of Subspecialty Medicine, Memorial Sloan Kettering Cancer Center.

[4]Department of Biostatistics, Epidemiology, and Informatics,

Perelman School of Medicine, University of Pennsylvania.

[1]Department of Statistics and Data Science, The Wharton School, University of Pennsylvania.

email: tingye1@uw.edu

Summary: Section S1 contains a review of the standard IV and DID; Section S2 includes additional results for the instrumented DID; Section S3 contains technical proofs; Section S4 contains additional details on the application.

This paper has been submitted for consideration for publication in *Biometrics*

S1. Review of standard IV and DID

S1.1 *Review: standard IV*

The standard IV method (except for the calendar time IV) does not have a time component and is most commonly applied to cross-sectional studies (Hernan and Robins, 2020). This is probably why many studies using the standard IV such as Neuman et al. (2014) simply ignore the longitudinal structure that is intrinsic in the dataset. To better understand the assumptions and make connections to the instrumented DID method proposed in Section 2, we embed the standard IV method into our potential outcomes framework with the time component made explicit.

Identification of the ATE β_0 using Z as a standard IV assumes that the following conditions hold with probability 1:

ASSUMPTION S1 (Standard IV): (a) (Relevance) $E(D \mid Z = 1, \boldsymbol{X}) \neq E(D \mid Z = 0, \boldsymbol{X})$; (b) (Independence & exclusion restriction) $Z \perp (T, D_t^{(z)}, Y_t^{(d)}, t = 0, 1, z = 0, 1, d = 0, 1) \mid \boldsymbol{X}$; (c) (No unmeasured common effect modifier) $\text{Cov}(D_t^{(1)} - D_t^{(0)}, Y_t^{(1)} - Y_t^{(0)} | \boldsymbol{X}) = 0$ for $t = 0, 1$.

Assumptions S1(a)-(b) formalize the three core IV assumptions discussed in Section 1. Assumption S1(a) says that Z is related to D. Assumption S1(b) says that Z is as good as random and has no direct effect on the outcome. In particular, $Z \perp T$ is required to guarantee that $(D_T^{(z)}, Y_T^{(d)}, z = 0, 1, d = 0, 1)$, the customarily defined potential exposures and outcomes when ignoring the longitudinal structure, are independent of Z. We refer interested readers to Richardson and Robins (2014) and Wang and Tchetgen Tchetgen (2018) for discussions on different statements of IV assumptions. Compared with the instrumented DID assumptions in Assumption 2, Assumption S1(a) is distinct from Assumption 2(a), Assumption S1(b) is stronger than Assumption 2(b), and Assumption S1(c) is the same as Assumption 2(c).

Under Assumption 1 and Assumption S1, we show in Section S3.1 that the conditional

Wald ratio identifies a weighted average of the time-specific CATE

$$\begin{aligned}&\frac{E(Y \mid Z=1,\boldsymbol{X})-E(Y \mid Z=0,\boldsymbol{X})}{E(D \mid Z=1,\boldsymbol{X})-E(D \mid Z=0,\boldsymbol{X})}\\&\quad=\frac{w_1(\boldsymbol{X})E(Y_1^{(1)}-Y_1^{(0)} \mid \boldsymbol{X})}{w_1(\boldsymbol{X})+w_0(\boldsymbol{X})}+\frac{w_0(\boldsymbol{X})E(Y_0^{(1)}-Y_0^{(0)} \mid \boldsymbol{X})}{w_1(\boldsymbol{X})+w_0(\boldsymbol{X})},\end{aligned} \tag{S1}$$

where $w_t(\boldsymbol{X})=P(T=t \mid \boldsymbol{X})E(D_t^{(1)}-D_t^{(0)} \mid \boldsymbol{X}), t=0,1$. This result indicates that if the treatment effect is expected to vary over time and the target estimand is the conditional or unconditional ATE, one should adjust for time indicator as an additional confounder.

S1.2 *Review: DID*

The basic DID framework considers a population observed in a pre-exposure period $t=0$ and a post-exposure period $t=1$. All individuals are unexposed at $t=0$. Some individuals become exposed at $t=1$. The parameter of interest is the average treatment effect for the treated in the post-exposure period $E(Y_1^{(1)}-Y_1^{(0)} \mid D_1=1)$, or the conditional counterpart $E(Y_1^{(1)}-Y_1^{(0)} \mid D_1=1,\boldsymbol{X})$.

The method of DID relies on a crucial parallel trends assumption $E(Y_1^{(0)}-Y_0^{(0)} \mid \boldsymbol{X}, D_1=1)=E(Y_1^{(0)}-Y_0^{(0)} \mid \boldsymbol{X}, D_1=0)$, which states that conditional on the covariates, the exposed and unexposed individuals would have exhibited parallel trends in the potential outcomes in the absence of treatment (Abadie, 2005). In other words, there is no unmeasured time-varying difference between the exposed and unexposed individuals. With this assumption, the effect of the treatment on the treated conditional on $\boldsymbol{X}$ is identified from

$$E(Y_1^{(1)}-Y_1^{(0)} \mid D_1=1,\boldsymbol{X})=E(Y_1-Y_0 \mid D_1=1,\boldsymbol{X})-E(Y_1-Y_0 \mid D_1=0,\boldsymbol{X}).$$

S2. Additional Results for Instrumented DID

S2.1 *Instrumented DID when treatment effect may change over time*

Consider the case without observed covariates. If Assumption 1 and Assumption 2(a)-(c) hold, then

$$\frac{\delta_Y}{\delta_D} = \frac{E(D_1^{(1)} - D_1^{(0)})}{\delta_D} E(Y_1^{(1)} - Y_1^{(0)}) - \frac{E(D_0^{(1)} - D_0^{(0)})}{\delta_D} E(Y_0^{(1)} - Y_0^{(0)}). \tag{S2}$$

When the treatment effect may vary over time, δ_Y/δ_D still has a nice interpretation under some special scenarios: (i) when either $E(D_0^{(1)} - D_0^{(0)})$ or $E(D_1^{(1)} - D_1^{(0)})$ is zero, then δ_Y/δ_D is the average treatment effect at the time point t in which $E(D_t^{(1)} - D_t^{(0)}) \neq 0$; (ii) when $E(D_0^{(1)} - D_0^{(0)})$ and $E(D_1^{(1)} - D_1^{(0)})$ are both non-zero and of opposite sign, then $E(D_0^{(1)} - D_0^{(0)})/\delta_D \in (-1, 0)$ and δ_Y/δ_D is a weighted average of $E(Y_1^{(1)} - Y_1^{(0)})$ and $E(Y_0^{(1)} - Y_0^{(0)})$ with non-negative weights. Otherwise, although δ_Y/δ_D is still a weighted average of the treatment effects at the two time points, the weights can be negative and δ_Y/δ_D no longer has a clear causal interpretation. For instance, if $E(Y_1^{(1)} - Y_1^{(0)}) > E(Y_0^{(1)} - Y_0^{(0)})$ and $E(D_1^{(1)} - D_1^{(0)}) > E(D_0^{(1)} - D_0^{(0)}) > 0$, then $\delta_Y/\delta_D > E(Y_1^{(1)} - Y_1^{(0)}) > E(Y_0^{(1)} - Y_0^{(0)})$, i.e., δ_Y/δ_D is larger than any time-specific average treatment effect.

S2.2 *One-sample and two-sample Wald estimators*

Let $\xrightarrow{d}$ denote convergence in distribution. Theorem S1 establishes the asymptotic property for the one-sample instrumented DID Wald estimator $\widehat{\beta}_{\text{wald}}$.

THEOREM S1: *Under Assumptions 1 and 2, and assume the second moments are finite, as $n \to \infty$, the Wald estimator $\widehat{\beta}_{\text{wald}}$ in (2) is consistent and asymptotically normal, i.e.,*

$$|\delta_D|\sqrt{n}(\widehat{\beta}_{\text{wald}} - \beta_0) \xrightarrow{d} N\left(0, \sum_{t=0,1}\sum_{z=0,1} \frac{Var(Y - \beta_0 D|T = t, Z = z)}{P(T = t, Z = z)}\right). \tag{S3}$$

For statistical inference, we can use a consistent plug-in variance estimator

$$\frac{1}{n(\widehat{\delta}_D)^2} \sum_{t=0,1}\sum_{z=0,1} \frac{\widehat{\text{Var}}(Y - \widehat{\beta}_{\text{wald}} D|T = t, Z = z)}{\widehat{P}(T = t, Z = z)}, \tag{S4}$$

where $\widehat{\delta}_D$ is defined in (2), $\widehat{P}(T=t,Z=z)=\mathbb{P}_n I_{(T=t,Z=z)}$, $\widehat{\mathrm{Var}}(Y-\widehat{\beta}_{\mathrm{wald}}D|T=t,Z=z)$ is the sample variance of $Y_i-\widehat{\beta}_{\mathrm{wald}}D_i$ within the stratum with $T_i=t,Z_i=z$.

The next theorem establishes the asymptotic property for the two-sample instrumented DID Wald estimator $\widehat{\beta}_{\mathrm{TSwald}}$.

THEOREM S2: *Suppose that Assumptions 1 and 2 hold for both* (T_a,Z_a,D_a,Y_a) *and* (T_b,Z_b,D_b,Y_b)*, and* $E(Y_a|T_a,Z_a)=E(Y_b|T_b,Z_b)$*,* $E(D_a|T_a,Z_a)=E(D_b|T_b,Z_b)$*. Also assume that* $\lim_{n_a,n_b\to\infty}\min(n_a,n_b)/n_c=\alpha_c\geqslant 0$ *for* $c\in\{a,b\}$*, and the second moments are finite. As* $\min(n_a,n_b)\to\infty$*, the two-sample Wald estimator* $\widehat{\beta}_{\mathrm{TSwald}}$ *is consistent and asymptotically normal, i.e.,*

$$|\delta_{Db}|\sqrt{\min(n_a,n_b)}(\widehat{\beta}_{\mathrm{TSwald}}-\beta_0)\xrightarrow{d}$$
$$N\left(0,\sum_{t=0,1}\sum_{z=0,1}\alpha_a\frac{Var(Y_a|T_a=t,Z_a=z)}{P(T_a=t,Z_a=z)}+\alpha_b\beta_0^2\frac{Var(D_b|T_b=t,Z_b=z)}{P(T_b=t,Z_b=z)}\right). \quad \text{(S5)}$$

For statistical inference, a consistent plug-in variance estimator for $\widehat{\beta}_{\mathrm{TSwald}}$ is

$$\frac{1}{(\widehat{\delta}_{Db})^2}\sum_{t=0,1}\sum_{z=0,1}\left[\widehat{\mathrm{Var}}\{\widehat{\mu}_{Ya}(t,z)\}+\widehat{\beta}^2_{\mathrm{TSwald}}\widehat{\mathrm{Var}}\{\widehat{\mu}_{Db}(t,z)\}\right], \quad \text{(S6)}$$

where $\widehat{\mu}_{Ya}(t,z)$ and $\widehat{\mu}_{Db}(t,z)$ are as defined in (2) but evaluated respectively at the outcome dataset and the exposure dataset, $\widehat{\mathrm{Var}}\{\widehat{\mu}_{Ya}(t,z)\}$ and $\widehat{\mathrm{Var}}\{\widehat{\mu}_{Db}(t,z)\}$ are their consistent variance estimators. In fact, $\widehat{\beta}_{\mathrm{TSwald}}$ and its variance estimator can be calculated provided that these summary statistics are available.

S2.3 *Sensitivity analysis*

We develop sensitivity analysis methods to evaluate how sensitive the conclusion is to violations of Assumption 2(d) for both the one-sample and two-sample designs when there are no observed covariates. There is a large and growing literature on sensitivity analysis, e.g., Rosenbaum (1987); Imbens (2003); VanderWeele and Ding (2017) and Fogarty (2020).

Consider first the one-sample design. When Assumption 2(d) does not hold, i.e., $E(Y_1^{(1)}-$

$Y_1^{(0)}) \neq E(Y_0^{(1)} - Y_0^{(0)})$. We use two sensitivity parameters γ_L, γ_U to quantify deviate from Assumption 2(d): $\Gamma := E(Y_1^{(1)} - Y_1^{(0)}) - E(Y_0^{(1)} - Y_0^{(0)}) \in [\gamma_L, \gamma_U]$, where $\gamma_L \leqslant 0 \leqslant \gamma_U$. When $\gamma_L = \gamma_U$, it is the same as the case under Assumption 2(d). Next, we construct a confidence interval for $\beta^* = E(Y_0^{(1)} - Y_0^{(0)})$ when $\Gamma \in [\gamma_L, \gamma_U]$; similar approach can be developed for $E(Y_1^{(1)} - Y_1^{(0)})$.

From (S2), we know that $\beta^* = \delta_Y/\delta_D - \Gamma\{\mu_D(1,1) - \mu_D(1,0)\}/\delta_D$, whose sample analogue is defined as $\widehat{\beta}_{\text{SA}}(\Gamma) = \widehat{\delta}_Y/\widehat{\delta}_D - \Gamma\{\widehat{\mu}_D(1,1) - \widehat{\mu}_D(1,0)\}/\widehat{\delta}_D$. Similar to the proof of Theorem S1, the asymptotic distribution of $\widehat{\beta}_{\text{SA}}(\Gamma)$ is

$$|\delta_D|\sqrt{n}(\widehat{\beta}_{\text{SA}}(\Gamma) - \beta^*) \xrightarrow{d} N\left(0, \sum_{t=0,1}\sum_{z=0,1} \frac{\text{Var}(Y - (\beta^* + t\Gamma)D | T = t, Z = z)}{P(T = t, Z = z)}\right).$$

Denote a consistent variance estimator of $\widehat{\beta}_{\text{SA}}(\Gamma)$ as $\widehat{V}_{\text{SA}}(\Gamma)$, let $C_L(\Gamma) = \widehat{\beta}_{\text{SA}}(\Gamma) - 1.96\widehat{V}_{\text{SA}}(\Gamma)^{1/2}$ and $C_U(\Gamma) = \widehat{\beta}_{\text{SA}}(\Gamma) + 1.96\widehat{V}_{\text{SA}}(\Gamma)^{1/2}$, then $[C_L(\Gamma), C_U(\Gamma)]$ is an asymptotic 95% confidence interval for β^* at any given value of Γ. By applying the union method (Zhao et al., 2019), we have that $\left[\inf_{\Gamma\in[\gamma_L,\gamma_U]} C_L(\Gamma), \sup_{\Gamma\in[\gamma_L,\gamma_U]} C_U(\Gamma)\right]$ is an asymptotic confidence interval with at least 95% coverage for any $\Gamma \in [\gamma_L, \gamma_U]$.

The sensitivity analysis for the two-sample setting is analogous. Define $\widehat{\beta}_{\text{SA,TS}}(\Gamma) = \widehat{\delta}_{Ya}/\widehat{\delta}_{Db} - \Gamma\{\widehat{\mu}_{Db}(1,1) - \widehat{\mu}_{Db}(1,0)\}/\widehat{\delta}_{Db}$. Similar to the proof of Theorem S2, the asymptotic distribution of $\widehat{\beta}_{\text{SA,TS}}(\Gamma)$ is

$$|\delta_{Db}|\sqrt{\min(n_a, n_b)}(\widehat{\beta}_{\text{SA,TS}}(\Gamma) - \beta^*) \xrightarrow{d}$$
$$N\left(0, \sum_{t=0,1}\sum_{z=0,1} \alpha_a \frac{\text{Var}(Y_a | T_a = t, Z_a = z)}{P(T_a = t, Z_a = z)} + \alpha_b(\beta^* + t\Gamma)^2 \frac{\text{Var}(D_b | T_b = t, Z_b = z)}{P(T_b = t, Z_b = z)}\right).$$

The construction of the confidence interval follows from the same steps as the one-sample design.

S2.4 *Bounded, efficient and multiply robust estimation*

In this section, motivated from Wang and Tchetgen Tchetgen (2018), we propose bounded versions of the semiparametric estimators that are guaranteed to fall within the parameter space with a binary outcome.

Note that for a binary Y, the nuisance parameters in $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ are no longer variation independent. For example, because $\delta(\boldsymbol{x})\delta_D(\boldsymbol{x}) + b_Y(\boldsymbol{x}) + m_{YT}(\boldsymbol{x}) + m_{YZ}(\boldsymbol{x}) = \mu_Y(1,1,\boldsymbol{x}) \in [0,1]$, the values taken by the other nuisance parameters will restrict the values that $\delta(\boldsymbol{x})$ can take. To make sure that the nuisance parameters are variation independent of each other, our choice of the nuisance parameters is

$$\begin{aligned} &\pi(t,z,\boldsymbol{x};\boldsymbol{\gamma}), \delta(\boldsymbol{x};\boldsymbol{\alpha}), \delta_D(\boldsymbol{x};\boldsymbol{\theta}), \\ &\mathrm{OP}_{D1}(\boldsymbol{x};\boldsymbol{\xi}_1), \mathrm{OP}_{D2}(\boldsymbol{x};\boldsymbol{\xi}_2), \mathrm{OP}_{D3}(\boldsymbol{x};\boldsymbol{\xi}_3), \mathrm{OP}_{Y1}(\boldsymbol{x};\boldsymbol{\vartheta}_1), \mathrm{OP}_{Y2}(\boldsymbol{x};\boldsymbol{\vartheta}_2), \mathrm{OP}_{Y3}(\boldsymbol{x};\boldsymbol{\vartheta}_3), \end{aligned} \tag{S7}$$

where for $C \in \{Y, D\}$ (recall the definition $\mu_C(t,z,\boldsymbol{X}) = E(C \mid T=t, Z=z, \boldsymbol{X})$),

$$\begin{aligned} \mathrm{OP}_{C1}(\boldsymbol{x}) &= \frac{\mu_C(0,1,\boldsymbol{x})\mu_C(0,0,\boldsymbol{x})}{\{1-\mu_C(0,1,\boldsymbol{x})\}\{1-\mu_C(0,0,\boldsymbol{x})\}}, \\ \mathrm{OP}_{C2}(\boldsymbol{x}) &= \frac{\mu_C(1,1,\boldsymbol{x})\mu_C(1,0,\boldsymbol{x})}{\{1-\mu_C(1,1,\boldsymbol{x})\}\{1-\mu_C(1,0,\boldsymbol{x})\}}, \\ \mathrm{OP}_{C3}(\boldsymbol{x}) &= \frac{\{1+\mu_C(1,1,\boldsymbol{x})-\mu_C(1,0,\boldsymbol{x})\}\{1+\mu_C(0,1,\boldsymbol{x})-\mu_C(0,0,\boldsymbol{x})\}}{\{1-\mu_C(1,1,\boldsymbol{x})+\mu_C(1,0,\boldsymbol{x})\}\{1-\mu_C(0,1,\boldsymbol{x})+\mu_C(0,0,\boldsymbol{x})\}}. \end{aligned}$$

Proposition S1 shows that our models provide a variation-independent parameterization of the likelihood $(P(Y=1 \mid T,Z,\boldsymbol{X}=\boldsymbol{x}), P(D=1 \mid T,Z,\boldsymbol{X}=\boldsymbol{x}))$. Also because $P(T,Z \mid \boldsymbol{X}=\boldsymbol{x})$ is variation-independent of $(P(Y=1 \mid T,Z,\boldsymbol{X}=\boldsymbol{x}), P(D=1 \mid T,Z,\boldsymbol{X}=\boldsymbol{x}))$, the parameter space of $(\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\boldsymbol{\xi}_3,\boldsymbol{\vartheta}_1,\boldsymbol{\vartheta}_2,\boldsymbol{\vartheta}_3)$ is unconstrained.

PROPOSITION S1: For binary D, Y and for any $\boldsymbol{x}$, the mapping

$$\begin{aligned} &(\delta(\boldsymbol{x}), \delta_D(\boldsymbol{x}), \mathrm{OP}_{D1}(\boldsymbol{x}), \mathrm{OP}_{D2}(\boldsymbol{x}), \mathrm{OP}_{D3}(\boldsymbol{x}), \mathrm{OP}_{Y1}(\boldsymbol{x}), \mathrm{OP}_{Y2}(\boldsymbol{x}), \mathrm{OP}_{Y3}(\boldsymbol{x})) \\ &\quad \to (\mu_D(t,z,\boldsymbol{x}), \mu_Y(t,z,\boldsymbol{x}), t=0,1, z=0,1) \end{aligned}$$

is a diffeomorphism between the interiors of their domains, which are $(-1,1)\times(-2,2)\times(0,\infty)^6$ and $(0,1)^8$ respectively.

Proof. We proceed with the proof in the following four steps:

(i) $\delta_Y(\boldsymbol{x}) = \delta_D(\boldsymbol{x})\delta(\boldsymbol{x}) \in (-2, 2)$;

(ii) The mapping $(\delta_D(\boldsymbol{x}), \mathrm{OP}_{D1}(\boldsymbol{x}), \mathrm{OP}_{D2}(\boldsymbol{x}), \mathrm{OP}_{D3}(\boldsymbol{x})) \to (\mu_D(t, z, \boldsymbol{x}), t = 0, 1, z = 0, 1)$ is a diffeomorphism from $(-2, 2) \times (0, \infty)^3$ to $(0, 1)^4$;

(iii) The mapping $(\delta_Y(\boldsymbol{x}), \mathrm{OP}_{Y1}(\boldsymbol{x}), \mathrm{OP}_{Y2}(\boldsymbol{x}), \mathrm{OP}_{Y3}(\boldsymbol{x})) \to (\mu_Y(t, z, \boldsymbol{x}), t = 0, 1, z = 0, 1)$ is a diffeomorphism from $(-2, 2) \times (0, \infty)^3$ to $(0, 1)^4$;

(iv) $P(Y = 1 \mid T, Z, \boldsymbol{X})$ is variation independent of $P(D = 1 \mid T, Z, \boldsymbol{X})$.

In the following, we show the second step. The third step can be shown in the same way. Note first that following a similar argument as in Richardson et al. (2017), $(\delta_D(\boldsymbol{x}), \mathrm{OP}_{D3}(\boldsymbol{x})) \to (\mu_D(0, 1, \boldsymbol{x}) - \mu_D(0, 0, \boldsymbol{x}), \mu_D(1, 1, \boldsymbol{x}) - \mu_D(1, 0, \boldsymbol{x}))$ is a diffeomorphism from $(-2, 2) \times (0, \infty)$ to $(-1, 1)^2$. Moreover, from Richardson et al. (2017), we have that $(\mathrm{OP}_{D1}(\boldsymbol{x}), \mu_D(0, 1, \boldsymbol{x}) - \mu_D(0, 0, \boldsymbol{x})) \to (\mu_D(0, 1, \boldsymbol{x}), \mu_D(0, 0, \boldsymbol{x}))$ is a diffeomorphism from $(0, \infty) \times (-1, 1)$ to $(0, 1)^2$, and $(\mathrm{OP}_{D2}(\boldsymbol{x}), \mu_D(1, 1, \boldsymbol{x}) - \mu_D(1, 0, \boldsymbol{x})) \to (\mu_D(1, 1, \boldsymbol{x}), \mu_D(1, 0, \boldsymbol{x}))$ is a diffeomorphism from $(0, \infty) \times (-1, 1)$ to $(0, 1)^2$. The result then follows from noting that $P(D = 1 \mid Z, T = 0, \boldsymbol{X})$ is variation independent of $P(D = 1 \mid Z, T = 1, \boldsymbol{X})$.

With the new set of nuisance parameters, we define $\boldsymbol{\Delta}^{\mathrm{b}}_C(\boldsymbol{x}) = (\mathrm{OP}_{C1}(\boldsymbol{x}), \mathrm{OP}_{C2}(\boldsymbol{x}), \mathrm{OP}_{C3}(\boldsymbol{x}))$, for $C \in \{Y, D\}$. Parallel to the development in Section 3.2, consider three sets of model assumptions:

$\mathcal{M}_1$: models for $\delta(\boldsymbol{x}), \delta_D(\boldsymbol{x}), \boldsymbol{\Delta}^{\mathrm{b}}_D(\boldsymbol{x}), \boldsymbol{\Delta}^{\mathrm{b}}_Y(\boldsymbol{x})$ are correct.

$\mathcal{M}_2$: models for $\pi(t, z, \boldsymbol{x}), \delta_D(\boldsymbol{x})$ are correct.

$\mathcal{M}_3$: models for $\pi(t, z, \boldsymbol{x}), \delta(\boldsymbol{x})$ are correct.

In what follows, we first present estimators of $\boldsymbol{\psi}_0$ under each of the three models. Importantly, we consider a working model $\beta(\boldsymbol{v}; \boldsymbol{\psi})$ that falls within the parameter space of $E(Y^{(1)} - Y^{(0)} \mid \boldsymbol{V} = \boldsymbol{v})$.

We first discuss regression-based estimation of $\boldsymbol{\psi}_0$ under model $\mathcal{M}_1$. For binary Y, we impose a model $\delta(\boldsymbol{x};\boldsymbol{\alpha})$ which guarantees that $\delta(\boldsymbol{x};\boldsymbol{\alpha}) \in [-1,1]$, e.g.,

$$\delta(\boldsymbol{x};\boldsymbol{\alpha}) = \tanh(\boldsymbol{\alpha}^T\boldsymbol{x}) = \frac{\exp(2\boldsymbol{\alpha}^T\boldsymbol{x})-1}{\exp(2\boldsymbol{\alpha}^T\boldsymbol{x})+1} \tag{S8}$$

as in Wang and Tchetgen Tchetgen (2018). For implementation, we first obtain the maximum likelihood estimators $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2, \widehat{\boldsymbol{\xi}}_3)$ from solving the score function

$$\mathbb{P}_n S(D \mid Z, T, \boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) = 0$$

corresponding to the likelihood of D conditional on $Z, T, \boldsymbol{X}$, then we obtain the maximum likelihood estimators $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\vartheta}}_1, \widehat{\boldsymbol{\vartheta}}_2, \widehat{\boldsymbol{\vartheta}}_3)$ from solving the score function

$$\mathbb{P}_n S(Y \mid Z, T, \boldsymbol{X}; \widehat{\boldsymbol{\theta}}, \boldsymbol{\alpha}, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \boldsymbol{\vartheta}_3) = 0$$

corresponding to the likelihood of Y conditional on $Z, T, \boldsymbol{X}$. The bounded regression-based estimator of $\boldsymbol{\psi}$ is the solution to

$$\mathbb{P}_n q(\boldsymbol{V};\boldsymbol{\psi})\{\delta(\boldsymbol{X};\widehat{\boldsymbol{\alpha}}) - \beta(\boldsymbol{V};\boldsymbol{\psi})\} = 0.$$

Under $\mathcal{M}_2$, the inverse probability weighting estimator is the same as that introduced in the main text, with the only difference of using a bounded working model $\beta(\boldsymbol{v};\boldsymbol{\psi})$.

Under $\mathcal{M}_3$, the estimator based on g-estimation is also the same as that introduced in the main text, with the only difference of using a bounded working model $\beta(\boldsymbol{v};\boldsymbol{\psi})$ and a bounded model for $\delta(\boldsymbol{x})$ as in (S8).

A bounded, efficient and multiply robust estimator can also be obtained from specifying nuisance models listed in (S7). As shown in Proposition S1, these models imply $b_D(\boldsymbol{x}), m_{DZ}(\boldsymbol{x}),\ m_{DT}(\boldsymbol{x}), b_Y(\boldsymbol{x}), m_{YZ}(\boldsymbol{x}), m_{YT}(\boldsymbol{x})$. Hence, a multiply robust estimator of ψ can be obtained in the same way introduced in main text, with $\mathrm{OP}_{D1}(\boldsymbol{x}), \mathrm{OP}_{D2}(\boldsymbol{x}), \mathrm{OP}_{D3}(\boldsymbol{x})$, $\mathrm{OP}_{Y1}(\boldsymbol{x}), \mathrm{OP}_{Y2}(\boldsymbol{x}), \mathrm{OP}_{Y3}(\boldsymbol{x})$ in place of $b_D(\boldsymbol{x}), m_{DZ}(\boldsymbol{x}),\ m_{DT}(\boldsymbol{x}), b_Y(\boldsymbol{x}), m_{YZ}(\boldsymbol{x}), m_{YT}(\boldsymbol{x})$.

S3. Technical Proofs

S3.1 *Proof of* (S1)

Note that from the property of conditional independence (Dawid, 1979, Lemma 4.3), Assumption 1(c) and Assumption S1(b) imply that $T \perp \{D_t^{(z)}, Y_t^{(d)}, t = 0, 1, z = 0, 1, d = 0, 1\} \mid \boldsymbol{X}$. Thus, the denominator in the Wald ratio in (S1) equals

$$
\begin{aligned}
&E[D_T^{(1)} \mid Z = 1, \boldsymbol{X}] - E[D_T^{(0)} \mid Z = 0, \boldsymbol{X}] \\
&= E[TD_1^{(1)} + (1-T)D_0^{(1)} \mid Z = 1, \boldsymbol{X}] - E[TD_1^{(0)} + (1-T)D_0^{(0)} \mid Z = 0, \boldsymbol{X}] \\
&= E(T \mid \boldsymbol{X})E[D_1^{(1)} \mid \boldsymbol{X}] + E(1-T \mid \boldsymbol{X})E[D_0^{(1)} \mid \boldsymbol{X}] \\
&\quad - E(T \mid \boldsymbol{X})E[D_1^{(0)} \mid \boldsymbol{X}] - E(1-T \mid \boldsymbol{X})E[D_0^{(0)} \mid \boldsymbol{X}] \\
&= E(T \mid \boldsymbol{X})E[D_1^{(1)} - D_1^{(0)} \mid \boldsymbol{X}] + E(1-T \mid \boldsymbol{X})E[D_0^{(1)} - D_0^{(0)} \mid \boldsymbol{X}] \\
&= E[T(D_1^{(1)} - D_1^{(0)}) \mid \boldsymbol{X}] + E[(1-T)(D_0^{(1)} - D_0^{(0)}) \mid \boldsymbol{X}].
\end{aligned}
$$

Similarly, the numerator in the Wald ratio in (S1) equals

$$
\begin{aligned}
&E[Y_T^{(D)}|Z = 1, \boldsymbol{X}] - E[Y_T^{(D)}|Z = 0, \boldsymbol{X}] \\
&= E[D_T^{(1)}Y_T^{(1)} + (1 - D_T^{(1)})Y_T^{(0)}|Z = 1, \boldsymbol{X}] - E[D_T^{(0)}Y_T^{(1)} + (1 - D_T^{(0)})Y_T^{(0)}|Z = 0, \boldsymbol{X}] \\
&= E[D_T^{(1)}Y_T^{(1)} + (1 - D_T^{(1)})Y_T^{(0)} \mid \boldsymbol{X}] - E[D_T^{(0)}Y_T^{(1)} + (1 - D_T^{(0)})Y_T^{(0)} \mid \boldsymbol{X}] \\
&= E[(D_T^{(1)} - D_T^{(0)})Y_T^{(1)} - (D_T^{(1)} - D_T^{(0)})Y_T^{(0)} \mid \boldsymbol{X}] \\
&= E[(D_T^{(1)} - D_T^{(0)})(Y_T^{(1)} - Y_T^{(0)}) \mid \boldsymbol{X}] \\
&= E[T(D_1^{(1)} - D_1^{(0)})(Y_1^{(1)} - Y_1^{(0)}) \mid \boldsymbol{X}] + E[(1-T)(D_0^{(1)} - D_0^{(0)})(Y_0^{(1)} - Y_0^{(0)}) \mid \boldsymbol{X}] \\
&= E[T(D_1^{(1)} - D_1^{(0)}) \mid \boldsymbol{X}]E[Y_1^{(1)} - Y_1^{(0)} \mid \boldsymbol{X}] + E[(1-T)(D_0^{(1)} - D_0^{(0)}) \mid \boldsymbol{X}]E[Y_0^{(1)} - Y_0^{(0)} \mid \boldsymbol{X}].
\end{aligned}
$$

This completes the proof.

S3.2 *Proof of Proposition 1*

Proof. First, note that for $z = 0, 1$,

$$\mu_Y(1, z, \boldsymbol{X}) - \mu_Y(0, z, \boldsymbol{X})$$

$$
\begin{aligned}
&E(Y|\boldsymbol{X}, T=1, Z=z) - E(Y|\boldsymbol{X}, T=0, Z=z) \\
&= E(Y_1^{(D_1^{(z)})}|\boldsymbol{X}, T=1, Z=z) - E(Y_0^{(D_0^{(z)})}|\boldsymbol{X}, T=0, Z=z) \\
&= E(Y_1^{(D_1^{(z)})} - Y_0^{(D_0^{(z)})}|\boldsymbol{X}, Z=z) \\
&= E(D_1^{(z)}Y_1^{(1)} + (1-D_1^{(z)})Y_1^{(0)} - D_0^{(z)}Y_0^{(1)} - (1-D_0^{(z)})Y_0^{(0)}|\boldsymbol{X}, Z=z) \\
&= E(D_1^{(z)}(Y_1^{(1)} - Y_1^{(0)}) - D_0^{(z)}(Y_0^{(1)} - Y_0^{(0)}) + Y_1^{(0)} - Y_0^{(0)}|\boldsymbol{X}, Z=z) \\
&= E(D_1^{(z)}(Y_1^{(1)} - Y_1^{(0)}) - D_0^{(z)}(Y_0^{(1)} - Y_0^{(0)})|\boldsymbol{X}) + E(Y_1^{(0)} - Y_0^{(0)}|\boldsymbol{X})
\end{aligned}
$$

where the second line is from Assumption 1(a), the third line is from Assumption 1(c), the last line is from Assumption 2(b). Thus,

$$
\begin{aligned}
\delta_Y(\boldsymbol{X}) &= E((D_1^{(1)} - D_1^{(0)})(Y_1^{(1)} - Y_1^{(0)})|\boldsymbol{X}) - E((D_0^{(1)} - D_0^{(0)})(Y_0^{(1)} - Y_0^{(0)})|\boldsymbol{X}) \\
&= E(D_1^{(1)} - D_1^{(0)}|\boldsymbol{X})E(Y_1^{(1)} - Y_1^{(0)}|\boldsymbol{X}) - E(D_0^{(1)} - D_0^{(0)}|\boldsymbol{X})E(Y_0^{(1)} - Y_0^{(0)}|\boldsymbol{X}) \\
&= E(D_1^{(1)} - D_1^{(0)} - D_0^{(1)} + D_0^{(0)}|\boldsymbol{X})\beta_0(\boldsymbol{X}) \\
&= \delta_D(\boldsymbol{X})\beta_0(\boldsymbol{X}),
\end{aligned}
$$

where the second line is from Assumption 2(c), the third line is from Assumption 2(d), the last line again uses Assumption 1(a)-(b).

S3.3 *Derivation of* $\delta_Y(\boldsymbol{X})/\delta_D(\boldsymbol{X})$ *under the monotonicity assumption*

From the proof of Proposition 1 and under the monotonicity assumption stated in the main article, we have

$$
\begin{aligned}
\delta_Y(\boldsymbol{X}) &= E((D_1^{(1)} - D_1^{(0)})(Y_1^{(1)} - Y_1^{(0)})|\boldsymbol{X}) - E((D_0^{(1)} - D_0^{(0)})(Y_0^{(1)} - Y_0^{(0)})|\boldsymbol{X}) \\
&= E(Y_1^{(1)} - Y_1^{(0)} \mid D_1^{(1)} - D_1^{(0)} = 1, \boldsymbol{X})P(D_1^{(1)} - D_1^{(0)} = 1|\boldsymbol{X}) \\
&\quad - E(Y_0^{(1)} - Y_0^{(0)} \mid D_0^{(1)} - D_0^{(0)} = 1|\boldsymbol{X})P(D_0^{(1)} - D_0^{(0)} = 1|\boldsymbol{X}) \\
&= E(Y_t^{(1)} - Y_t^{(0)} \mid D_t^{(1)} - D_t^{(0)} = 1)\left\{P(D_1^{(1)} - D_1^{(0)} = 1|\boldsymbol{X}) - P(D_0^{(1)} - D_0^{(0)} = 1|\boldsymbol{X})\right\},
\end{aligned}
$$

where the last line is from the assumption that $E(Y_1^{(1)} - Y_1^{(0)} \mid D_1^{(1)} - D_1^{(0)} = 1) = E(Y_0^{(1)} - Y_0^{(0)} \mid D_0^{(1)} - D_0^{(0)} = 1)$. In addition, $\delta_D(\boldsymbol{X}) = P(D_1^{(1)} - D_1^{(0)} = 1|\boldsymbol{X}) - P(D_0^{(1)} - D_0^{(0)} = 1|\boldsymbol{X})$. This completes the proof.

S3.4 *Proof of Theorem S1*

From the definition of $\widehat{\beta}$,

$$\sqrt{n}(\widehat{\beta} - \beta_0) = \frac{\sqrt{n}(\widehat{\delta}_Y - \beta_0\widehat{\delta}_D)}{\widehat{\delta}_D}.$$

Let $\mathcal{F} = \{T_i, Z_i, i = 1, \ldots, n\}$ and

$$K_i = \sqrt{n}(Y_i - \beta_0 D_i)\Bigg\{ \frac{I(T_i = 1, Z_i = 1)}{\sum_{i=1}^n I(T_i = 1, Z_i = 1)} - \frac{I(T_i = 1, Z_i = 0)}{\sum_{i=1}^n I(T_i = 1, Z_i = 0)} - \frac{I(T_i = 0, Z_i = 1)}{\sum_{i=1}^n I(T_i = 0, Z_i = 1)} + \frac{I(T_i = 0, Z_i = 0)}{\sum_{i=1}^n I(T_i = 0, Z_i = 0)} \Bigg\}.$$

Then, we can write

$$\sqrt{n}(\widehat{\delta}_Y - \beta_0\widehat{\delta}_D) = \sum_{i=1}^n K_i.$$

First, note that K_i, $i = 1, \ldots, n$ are independent conditional on $\mathcal{F}$, and $E(\sum_{i=1}^n K_i|\mathcal{F}) = \sqrt{n}(\delta_Y - \beta_0\delta_D) = 0$, and

$$\mathrm{Var}(K_i|\mathcal{F}) = n\sum_{t=0}^1\sum_{z=0}^1 \mathrm{Var}(Y - \beta_0 D|T = t, Z = z)\frac{I(T_i = t, Z_i = z)}{\{\sum_{i=1}^n I(T_i = t, Z_i = z)\}^2}.$$

We prove that $\sum_{i=1}^n K_i$ is asymptotically normal by verifying Lindeberg's condition. Let

$$\sigma^2 = \sum_{i=1}^n \mathrm{Var}(K_i|\mathcal{F}) = \sum_{t=0}^1\sum_{z=0}^1 \frac{\mathrm{Var}(Y - \beta_0 D|T = t, Z = z)}{n^{-1}\sum_{i=1}^n I(T_i = t, Z_i = z)},$$

we have that

$$\frac{\max_i \mathrm{Var}(K_i|\mathcal{F})}{\sigma^2} = \max_{t',z'} \frac{\frac{\mathrm{Var}(Y-\beta_0 D|T=t',Z=z')}{\{\sum_{i=1}^n I(T_i=t',Z_i=z')\}^2}}{\sum_{z=0}^1\sum_{z=0}^1 \frac{\mathrm{Var}(Y-\beta_0 D|T=t,Z=z)}{\sum_{i=1}^n I(T_i=t,Z_i=z)}} \leqslant \max_{t',z'} \frac{\frac{\mathrm{Var}(Y-\beta_0 D|T=t',Z=z')}{\{\sum_{i=1}^n I(T_i=t',Z_i=z')\}^2}}{\frac{\mathrm{Var}(Y-\beta_0 D|T=t',Z=z')}{\sum_{i=1}^n I(T_i=t',Z_i=z')}}$$

$$= \max_{t',z'} \frac{1}{\sum_{i=1}^n I(T_i = t', Z_i = z')} = o(1).$$

Hence, for any $\epsilon > 0$,

$$\sum_{i=1}^n E\left\{ \frac{(K_i - E(K_i|\mathcal{F}))^2}{\sigma^2} I\left(|K_i - E(K_i|\mathcal{F})| > \epsilon\sigma\right) \mid \mathcal{F} \right\}$$

$$= \sum_{i=1}^n \frac{\mathrm{Var}(K_i|\mathcal{F})}{\sigma^2} E\left\{ \frac{(K_i - E(K_i|\mathcal{F}))^2}{\mathrm{Var}(K_i|\mathcal{F})} I\left(|K_i - E(K_i|\mathcal{F})| > \epsilon\sigma\right) \mid \mathcal{F} \right\}$$

$$\leqslant \max_i E\left\{\frac{(K_i - E(K_i|\mathcal{F}))^2}{\mathrm{Var}(K_i|\mathcal{F})} I\left(\frac{|K_i - E(K_i|\mathcal{F})|}{\sqrt{\mathrm{Var}(K_i|\mathcal{F})}} > \frac{\epsilon\sigma}{\sqrt{\mathrm{Var}(K_i|\mathcal{F})}}\right) \mid \mathcal{F}\right\}$$

$$= o(1),$$

where the last equality is from dominated convergence theorem and the facts that $\{K_i - E(K_i|\mathcal{F})\}/\sqrt{\mathrm{Var}(K_i|\mathcal{F})}$ has expectation zero and variance 1 conditional on $\mathcal{F}$, and $\max_i \mathrm{Var}(K_i|\mathcal{F})/\sigma^2 = o(1)$. Therefore, Lindeberg's condition holds. Applying Linderberg Central Limit Theorem, we have that conditional on $\mathcal{F}$,

$$\frac{\sqrt{n}(\widehat{\delta}_Y - \beta_0\widehat{\delta}_D)}{\sigma} \mid \mathcal{F} \xrightarrow{d} N(0,1).$$

By a dominated convergence argument, we have that the above equation also holds unconditionally. Then, by weak law of large numbers and Slutsky's theorem, it is easy to show that

$$\sigma^2 = \sum_{t=0}^{1}\sum_{z=0}^{1} \frac{\mathrm{Var}(Y - \beta_0 D|T=t, Z=z)}{P(T=t, Z=z)} + o_p(1),$$

and

$$\sqrt{n}(\widehat{\delta}_Y - \beta_0\widehat{\delta}_D) \xrightarrow{d} N\left(0, \sum_{t=0}^{1}\sum_{z=0}^{1} \frac{\mathrm{Var}(Y - \beta_0 D|T=t, Z=z)}{P(T=t, Z=z)}\right).$$

Finally, we can similarly show that $\sqrt{n}(\widehat{\delta}_D - \delta_D)$ is asymptotically normal, which implies that $\widehat{\delta}_D \xrightarrow{p} \delta_D$. Again using Slutsky's theorem, we have proved (S3).

S3.5 *Proof of Theorem 1*

In this section, we use subscripts to explicitly index quantities that depend on the distribution P, we use a zero subscript to denote a quantity evaluated at the true distribution $P = P_0$, we use a ϵ subscript to denote a quantity evaluated at the parametric submodel $P = P_\epsilon$. We will show that $\varphi(O; \boldsymbol{\psi}_P, \boldsymbol{\eta}_P)$ is proportional to the efficient influence function by showing that it is the canonical gradient of the pathwise derivative of $\boldsymbol{\psi}_P$, i.e,

$$\left.\frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon}\right|_{\epsilon=0} = C_0^{-1} E_0\left\{\varphi(O; \boldsymbol{\psi}_P, \boldsymbol{\eta}_P) s_0(\boldsymbol{O})\right\}, \tag{S9}$$

where $\boldsymbol{\psi}_\epsilon = \boldsymbol{\psi}_{P_\epsilon}$, $s_\epsilon(\boldsymbol{O}) = \partial \log dP_\epsilon(\boldsymbol{O})/\partial\epsilon$ denotes the parameter submodel score, C_0 is defined later in (S10).

By definition, we have

$$\boldsymbol{\psi}_P = \arg\min_{\boldsymbol{\psi}} \int w(\boldsymbol{v})\{\beta_P(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi})\}^2 dP(\boldsymbol{v}),$$

and thus

$$\int q(\boldsymbol{v};\boldsymbol{\psi})\{\beta_P(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi})\} dP(\boldsymbol{v}) = 0,$$

where $q(\boldsymbol{v};\boldsymbol{\psi}) = w(\boldsymbol{v})\frac{\partial\beta(\boldsymbol{v};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}$. Evaluating the above at $P = P_\epsilon$ gives

$$\int q(\boldsymbol{v};\boldsymbol{\psi}_\epsilon)\{\beta_\epsilon(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_\epsilon)\} dP_\epsilon(\boldsymbol{v}) = 0,$$

Differentiating the above with respect to ϵ using the chain rule and evaluating at the truth $\epsilon = 0$ give

$$\begin{aligned}
&\int \frac{\partial q(\boldsymbol{v};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \frac{\partial\boldsymbol{\psi}_\epsilon}{\partial\epsilon}\bigg|_{\epsilon=0} \{\beta_0(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_0)\}\, dP_0(\boldsymbol{v}) \\
&+ \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \left\{ \frac{\partial\beta_\epsilon(\boldsymbol{v})}{\partial\epsilon}\bigg|_{\epsilon=0} - \frac{\partial\beta(\boldsymbol{v};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \frac{\partial\boldsymbol{\psi}_\epsilon}{\partial\epsilon}\bigg|_{\epsilon=0} \right\} dP_0(\boldsymbol{v}) \\
&+ \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \{\beta_0(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_0)\} s_0(\boldsymbol{v}) dP_0(\boldsymbol{v}) = 0.
\end{aligned}$$

Rearranging the above equation, we have

$$\begin{aligned}
&\frac{\partial\boldsymbol{\psi}_\epsilon}{\partial\epsilon}\bigg|_{\epsilon=0} \underbrace{\int \left[\frac{\partial q(\boldsymbol{v};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \{\beta_0(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_0)\} - q(\boldsymbol{v};\boldsymbol{\psi}_0) \frac{\partial\beta(\boldsymbol{v};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}_0} \right] dP_0(\boldsymbol{v})}_{-C_0} \qquad \text{(S10)} \\
&+ \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \left\{ \frac{\partial\beta_\epsilon(\boldsymbol{v})}{\partial\epsilon}\bigg|_{\epsilon=0} + \{\beta_0(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_0)\} s_0(\boldsymbol{v}) \right\} dP_0(\boldsymbol{v}) = 0,
\end{aligned}$$

and thus

$$C_0 \frac{\partial\boldsymbol{\psi}_\epsilon}{\partial\epsilon}\bigg|_{\epsilon=0} = \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \left\{ \frac{\partial\beta_\epsilon(\boldsymbol{v})}{\partial\epsilon}\bigg|_{\epsilon=0} + \{\beta_0(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_0)\} s_0(\boldsymbol{v}) \right\} dP_0(\boldsymbol{v}).$$

Next, we will derive $\frac{\partial\beta_\epsilon(\boldsymbol{v})}{\partial\epsilon}|_{\epsilon=0}$. Note that

$$\begin{aligned}
&\frac{\partial\beta_\epsilon(\boldsymbol{v})}{\partial\epsilon}\bigg|_{\epsilon=0} \\
&= \frac{\partial}{\partial\epsilon} E_\epsilon \left[\frac{\delta_{Y\epsilon}(\boldsymbol{X})}{\delta_{D\epsilon}(\boldsymbol{X})} \bigg| \boldsymbol{V} = \boldsymbol{v} \right] \bigg|_{\epsilon=0}
\end{aligned}$$

$$= \frac{\partial}{\partial \epsilon} \int \frac{\delta_{Y\epsilon}(\boldsymbol{X})}{\delta_{D\epsilon}(\boldsymbol{X})} dP_\epsilon(\boldsymbol{X}|\boldsymbol{V}=\boldsymbol{v})\Big|_{\epsilon=0}$$
$$= \int \left[\frac{\frac{\partial \delta_{Y\epsilon}(\boldsymbol{X})}{\partial \epsilon}|_{\epsilon=0} \delta_{D0}(\boldsymbol{X}) - \delta_{Y0}(\boldsymbol{X}) \frac{\partial \delta_{D\epsilon}(\boldsymbol{X})}{\partial \epsilon}|_{\epsilon=0}}{[\delta_{D0}(\boldsymbol{X})]^2} + \frac{\delta_{Y0}(\boldsymbol{X})}{\delta_{D0}(\boldsymbol{X})} s_0(\boldsymbol{X}|\boldsymbol{V}) \right] dP_0(\boldsymbol{X}|\boldsymbol{V}=\boldsymbol{v}),$$

and

$$\frac{\partial \delta_{Y\epsilon}(\boldsymbol{X})}{\partial \epsilon}\Big|_{\epsilon=0}$$
$$= E_0[Y s_0(Y|T,Z,\boldsymbol{X})|T=1,Z=1,\boldsymbol{X}] - E_0[Y s_0(Y|T,Z,\boldsymbol{X})|T=0,Z=1,\boldsymbol{X]}$$
$$- E_0[Y s_0(Y|T,Z,\boldsymbol{X})|T=1,Z=0,\boldsymbol{X}] + E_0[Y s_0(Y|T,Z,\boldsymbol{X})|T=0,Z=0,\boldsymbol{X}]$$
$$= E_0 \left[\left\{ \frac{TZ}{P_0(T=1,Z=1|\boldsymbol{X})} - \frac{(1-T)Z}{P_0(T=0,Z=1|\boldsymbol{X})} \right.\right.$$
$$\left.\left. - \frac{T(1-Z)}{P_0(T=1,Z=0|\boldsymbol{X})} + \frac{(1-T)(1-Z)}{P_0(T=0,Z=0|\boldsymbol{X})} \right\} Y s_0(Y|T,Z,\boldsymbol{X}) \Big| \boldsymbol{X} \right]$$
$$= E_0 \left[\frac{(2Z-1)(2T-1)}{\pi_0(T,Z,\boldsymbol{X})} Y s_0(Y|T,Z,\boldsymbol{X}) | \boldsymbol{X} \right],$$

where $\pi_0(t,z,\boldsymbol{X}) = P_0(T=t,Z=z|\boldsymbol{X})$. Similarly, we can also derive that

$$\frac{\partial \delta_{D\epsilon}(X)}{\partial \epsilon}\Big|_{\epsilon=0} = E_0 \left[\frac{(2Z-1)(2T-1)}{\pi_0(T,Z,\boldsymbol{X})} D s_0(D|T,Z,\boldsymbol{X}) | \boldsymbol{X} \right].$$

Combining the above derivations, we have

$$C_0 \frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon}\Big|_{\epsilon=0}$$
$$= \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \frac{\partial \beta_\epsilon(\boldsymbol{v})}{\partial \epsilon}\Big|_{\epsilon=0} dP_0(\boldsymbol{v}) + \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \left\{ \beta_0(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_0) \right\} s_0(\boldsymbol{v}) dP_0(\boldsymbol{v})$$
$$= \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \left[\frac{\frac{\partial \delta_{Y\epsilon}(\boldsymbol{X})}{\partial \epsilon}|_{\epsilon=0} \delta_{D0}(\boldsymbol{X}) - \delta_{Y0}(\boldsymbol{X}) \frac{\partial \delta_{D\epsilon}(\boldsymbol{X})}{\partial \epsilon}|_{\epsilon=0}}{[\delta_{D0}(\boldsymbol{X})]^2} \right.$$
$$\left. + \frac{\delta_{Y0}(\boldsymbol{X})}{\delta_{D0}(\boldsymbol{X})} s_0(\boldsymbol{X}|\boldsymbol{V}) \right] dP_0(\boldsymbol{X}|\boldsymbol{V}=\boldsymbol{v}) dP_0(\boldsymbol{v})$$
$$+ \int q(\boldsymbol{v};\boldsymbol{\psi}_0) \left\{ \beta_0(\boldsymbol{v}) - \beta(\boldsymbol{v};\boldsymbol{\psi}_0) \right\} s_0(\boldsymbol{v}) dP_0(\boldsymbol{v}). \tag{S11}$$

We now turn to $E_0\{\varphi(\boldsymbol{O};\boldsymbol{\psi}_P,\boldsymbol{\eta}_P) s_0(\boldsymbol{O})\}$. First, note that $\varphi(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\eta})$ can be rewritten as

$$\varphi(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\eta}) = q(\boldsymbol{V};\boldsymbol{\psi}) \Big(\delta(\boldsymbol{X}) - \beta(\boldsymbol{V};\boldsymbol{\psi}) + \frac{(2Z-1)(2T-1)}{\pi(T,Z,\boldsymbol{X})\delta_D(\boldsymbol{X})}$$
$$\left[Y - E(Y \mid T,Z,\boldsymbol{X}) - \delta(\boldsymbol{X})\{D - E(D \mid T,Z,\boldsymbol{X})\} \right] \Big).$$

Also note that $s_0(\boldsymbol{O})$ is the parametric submodel score can be decomposed as

$$s_0(\boldsymbol{O}) = s_0(Y, D|T, Z, \boldsymbol{X}) + s_0(T, Z|\boldsymbol{X}) + s_0(\boldsymbol{X}|\boldsymbol{V}) + s_0(\boldsymbol{V}).$$

Then, with the scaling factor, the efficient influence function is $C_0^{-1}\varphi(\boldsymbol{O}; \boldsymbol{\psi}_P, \boldsymbol{\eta}_P)$, where $\varphi(\boldsymbol{O}; \boldsymbol{\psi}_P, \boldsymbol{\eta}_P)$ is defined in Theorem 1. Therefore,

$$\begin{aligned}
&E_0\{\varphi(\boldsymbol{O}; \boldsymbol{\psi}_0, \boldsymbol{\eta}_0)s_0(\boldsymbol{O})\} \\
&= E_0\left\{q(\boldsymbol{V}; \boldsymbol{\psi}_0)\left[\frac{\delta_{Y0}(\boldsymbol{X})}{\delta_{D0}(\boldsymbol{X})} - \beta(\boldsymbol{V}; \boldsymbol{\psi}_0)\right]\{s_0(\boldsymbol{X}|\boldsymbol{V}) + s_0(V)\}\right\} \\
&\quad + E_0\left\{q(\boldsymbol{V}; \boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T, Z, \boldsymbol{X})\delta_{D0}(\boldsymbol{X})}\left[Y - E_0(Y|T, Z, \boldsymbol{X})\right]s_0(Y|T, Z, \boldsymbol{X})\right\} \\
&\quad - E_0\left\{q(\boldsymbol{V}; \boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T, Z, \boldsymbol{X})\delta_{D0}(\boldsymbol{X})}\frac{\delta_{Y0}(\boldsymbol{X})}{\delta_{D0}(\boldsymbol{X})}[D - E_0(D|T, Z, \boldsymbol{X})]s_0(D|T, Z, \boldsymbol{X})\right\} \\
&= E_0\left\{q(\boldsymbol{V}; \boldsymbol{\psi}_0)\frac{\delta_{Y0}(\boldsymbol{X})}{\delta_{D0}(\boldsymbol{X})}s_0(\boldsymbol{X}|\boldsymbol{V})\right\} + E_0\left\{q(\boldsymbol{V}; \boldsymbol{\psi}_0)\left[\beta_0(\boldsymbol{V}) - \beta(\boldsymbol{V}; \boldsymbol{\psi}_0)\right]s_0(V)\right\} \\
&\quad + E_0\left\{q(\boldsymbol{V}; \boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T, Z, \boldsymbol{X})\delta_{D0}(\boldsymbol{X})}Y s_0(Y|T, Z, \boldsymbol{X})\right\} \\
&\quad - E_0\left\{q(\boldsymbol{V}; \boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T, Z, \boldsymbol{X})\delta_{D0}(\boldsymbol{X})}\frac{\delta_{Y0}(\boldsymbol{X})}{\delta_{D0}(\boldsymbol{X})}D s_0(D|T, Z, \boldsymbol{X})\right\} \\
&= C_0\frac{\partial \boldsymbol{\psi}_\epsilon}{\partial \epsilon}\bigg|_{\epsilon=0},
\end{aligned}$$

where the derivations follow from $E_0(s_0(\boldsymbol{O}_1|\boldsymbol{O}_2)|\boldsymbol{O}_2) = 0$ for any $(\boldsymbol{O}_1, \boldsymbol{O}_2) \subset \boldsymbol{O}$ and iterated expectation. Hence, $C_0^{-1}\varphi(\boldsymbol{O}; \boldsymbol{\psi}_P, \boldsymbol{\eta}_P)$ is the efficient influence function.

S3.6 *Proof of multiple robustness*

From the definition of $\boldsymbol{\psi}_0$ in (3), it is true that

$$E\left[q(\boldsymbol{V}; \boldsymbol{\psi}_0)\left\{\beta_0(\boldsymbol{V}) - \beta(\boldsymbol{V}; \boldsymbol{\psi}_0)\right\}\right] = 0. \tag{S12}$$

Under $\mathcal{M}_1$, $\overline{\delta}(\boldsymbol{X}) = \delta_0(\boldsymbol{X}), \overline{b}_Y(\boldsymbol{X}) = b_{Y0}(\boldsymbol{X}), \overline{b}_D(\boldsymbol{X}) = b_{D0}(\boldsymbol{X}), \overline{m}_{YZ}(\boldsymbol{X}) = m_{YZ0}(\boldsymbol{X}), \overline{m}_{YT}(\boldsymbol{X}) = m_{YT0}(\boldsymbol{X}), \overline{m}_{DZ}(\boldsymbol{X}) = m_{DZ0}(\boldsymbol{X}), \overline{m}_{DT}(\boldsymbol{X}) = m_{DT0}(\boldsymbol{X})$. Then,

$$\begin{aligned}
&E[\varphi(\boldsymbol{O}; \boldsymbol{\psi}_0, \overline{\boldsymbol{\eta}})] \\
&= E\left[q(\boldsymbol{V}; \boldsymbol{\psi}_0)\left\{\delta_0(\boldsymbol{X}) - \beta(\boldsymbol{V}; \boldsymbol{\psi}_0)\right\}\right] \\
&\quad + E\left[q(\boldsymbol{V}; \boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\overline{\pi}(T, Z, \boldsymbol{X})\overline{\delta}_D(\boldsymbol{X})}(Y - b_{Y0}(\boldsymbol{X}) - m_{YZ0}(\boldsymbol{X})Z - m_{YT0}(\boldsymbol{X})T)\right]
\end{aligned}$$

$$
\begin{aligned}
&\quad - E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\overline{\pi}(T,Z,\boldsymbol{X})\overline{\delta}_D(\boldsymbol{X})}\delta_0(\boldsymbol{X})(D-b_{D0}(\boldsymbol{X})-m_{DZ0}(\boldsymbol{X})Z-m_{DT0}(\boldsymbol{X})T)\right]\\
&= E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\left\{\delta_0(\boldsymbol{X})-\beta(\boldsymbol{V};\boldsymbol{\psi}_0)\right\}\right]+E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\overline{\pi}(T,Z,\boldsymbol{X})\overline{\delta}_D(\boldsymbol{X})}\delta_{D0}(\boldsymbol{X})\delta_0(\boldsymbol{X})TZ\right]\\
&\quad - E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\overline{\pi}(T,Z,\boldsymbol{X})\overline{\delta}_D(\boldsymbol{X})}\delta_{D0}(\boldsymbol{X})\delta_0(\boldsymbol{X})TZ\right]\\
&= E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\left\{\beta_0(\boldsymbol{V})-\beta(\boldsymbol{V};\boldsymbol{\psi}_0)\right\}\right]=0.
\end{aligned}
$$

Under $\mathcal{M}_2$, $\overline{\pi}(T,Z,\boldsymbol{X})=\pi_0(T,Z,\boldsymbol{X})$ and $\overline{\delta}_D(\boldsymbol{X})=\delta_{D0}(\boldsymbol{X})$. Then,

$$
\begin{aligned}
&E[\varphi(\boldsymbol{O};\boldsymbol{\psi}_0,\overline{\boldsymbol{\eta}})]\\
&= E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\left\{\overline{\delta}(\boldsymbol{X})-\beta(\boldsymbol{V};\boldsymbol{\psi}_0)\right\}\right]\\
&\quad + E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T,Z,\boldsymbol{X})\delta_{D0}(\boldsymbol{X})}(Y-\overline{b}_Y(\boldsymbol{X})-\overline{m}_{YZ}(\boldsymbol{X})Z-\overline{m}_{YT}(\boldsymbol{X})T)\right]\\
&\quad - E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T,Z,\boldsymbol{X})\delta_{D0}(\boldsymbol{X})}\overline{\delta}(\boldsymbol{X})(D-\overline{b}_D(\boldsymbol{X})-\overline{m}_{DZ}(\boldsymbol{X})Z-\overline{m}_{DT}(\boldsymbol{X})T)\right]\\
&= E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\left\{\overline{\delta}(\boldsymbol{X})-\beta(\boldsymbol{V};\boldsymbol{\psi}_0)\right\}\right]+E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\{\delta_0(\boldsymbol{X})-\overline{\delta}(\boldsymbol{X})\}\right]\\
&= E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\left\{\beta_0(\boldsymbol{V})-\beta(\boldsymbol{V};\boldsymbol{\psi}_0)\right\}\right]=0,
\end{aligned}
$$

where the third equality uses the facts that $E(Y|T,Z,\boldsymbol{X})=b_{Y0}(\boldsymbol{X})+m_{YZ0}(\boldsymbol{X})Z+m_{YT0}(\boldsymbol{X})T+\delta_0(\boldsymbol{X})\delta_{D0}(\boldsymbol{X})TZ$, $E(D|T,Z,\boldsymbol{X})=b_{D0}(\boldsymbol{X})+m_{DZ0}(\boldsymbol{X})Z+m_{DT0}(\boldsymbol{X})T+\delta_{D0}(\boldsymbol{X})TZ$, and $E\{(2Z-1)(2T-1)/\pi_0(T,Z,\boldsymbol{X})|T,\boldsymbol{X}\}=E\{(2Z-1)(2T-1)/\pi_0(T,Z,\boldsymbol{X})|Z,\boldsymbol{X}\}=0$. Hence, the efficient influence function $\varphi(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\eta})$ has expectation zero at $\boldsymbol{\psi}=\boldsymbol{\psi}_0$ under $\mathcal{M}_2$.

Under $\mathcal{M}_3$, $\overline{\pi}(T,Z,\boldsymbol{X})=\pi_0(T,Z,\boldsymbol{X})$, $\overline{\delta}(\boldsymbol{X})=\delta_0(\boldsymbol{X})$. Then,

$$
\begin{aligned}
&E[\varphi(\boldsymbol{O};\boldsymbol{\psi}_0,\overline{\boldsymbol{\eta}})]\\
&= E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\left\{\delta_0(\boldsymbol{X})-\beta(\boldsymbol{V};\boldsymbol{\psi}_0)\right\}\right]+E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T,Z,\boldsymbol{X})\overline{\delta}_D(\boldsymbol{X})}\delta_0(\boldsymbol{X})\delta_{D0}(\boldsymbol{X})TZ\right]\\
&\quad - E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\frac{(2Z-1)(2T-1)}{\pi_0(T,Z,\boldsymbol{X})\overline{\delta}_D(\boldsymbol{X})}\delta_0(\boldsymbol{X})\delta_{D0}(\boldsymbol{X})TZ\right]\\
&= E\left[q(\boldsymbol{V};\boldsymbol{\psi}_0)\left\{\delta_0(\boldsymbol{V})-\beta(\boldsymbol{V};\boldsymbol{\psi}_0)\right\}\right]=0.
\end{aligned}
$$

Hence, the efficient influence function $\varphi(\boldsymbol{O};\boldsymbol{\psi},\boldsymbol{\eta})$ has expectation zero at $\boldsymbol{\psi}=\boldsymbol{\psi}_0$ under $\mathcal{M}_3$.

S3.7 *Proof of Theorem 2*

In what follows, we will use $P\{f(\boldsymbol{O})\} = \int f(\boldsymbol{O})dP$ to denote expectation treating the function f as fixed; thus $P\{f(\boldsymbol{O})\}$ is random when f is random, and is different from the fixed quantity $E\{f(\boldsymbol{O})\}$ which averages over randomness in both f and $\boldsymbol{O}$.

Since $\widehat{\boldsymbol{\psi}}$ is a Z-estimator, using Theorem 5.31 of van der Vaart (2000), we have that under Assumption 3,

$$\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) = -M^{-1}_{\boldsymbol{\psi}_0,\overline{\boldsymbol{\eta}}}\sqrt{n}P\{\varphi(\boldsymbol{O};\boldsymbol{\psi}_0,\widehat{\boldsymbol{\eta}})\} - M^{-1}_{\boldsymbol{\psi}_0,\overline{\boldsymbol{\eta}}}n^{-1/2}\sum_{i=1}^{n}[\varphi(\boldsymbol{O}_i;\boldsymbol{\psi}_0,\overline{\boldsymbol{\eta}}) - E\{\varphi(\boldsymbol{O};\boldsymbol{\psi}_0,\overline{\boldsymbol{\eta}})\}]$$
$$+ o_p(1 + \sqrt{n}\|P\{\varphi(\boldsymbol{O};\boldsymbol{\psi}_0,\widehat{\boldsymbol{\eta}})\}\|),$$

Using standard central limit theorem, the second term is asymptotically normal, and is $O_p(1)$. Hence, the consistency and rate of convergence of $\widehat{\boldsymbol{\psi}}$ depends on the property of the first term. We analyze $\sqrt{n}P\{\varphi(\boldsymbol{O};\boldsymbol{\psi}_0,\widehat{\boldsymbol{\eta}})\}$ in the following.

For ease of exposition, we will simplify the notations to $q, \mu_Y, \mu_D, \delta_Y, \delta_D, \pi$ and keep the involved random variables implicit. Also let $\widehat{\Lambda}_Y = \widehat{b}_Y + \widehat{m}_{YZ}Z + \widehat{m}_{YT}T$, $\widehat{\Lambda}_D = \widehat{b}_D + \widehat{m}_{DZ}Z + \widehat{m}_{DT}T$, $\Lambda_{Y0} = b_{Y0} + m_{YZ0}Z + m_{YT0}T$, and $\Lambda_{D0} = b_{D0} + m_{DZ0}Z + m_{DT0}T$. Note that

$$\begin{aligned}
&P\{\varphi(\boldsymbol{O};\boldsymbol{\psi}_0,\widehat{\boldsymbol{\eta}})\}\\
&= P\left[q\{\widehat{\delta} - \delta_0 + \frac{(2Z-1)(2T-1)}{\widehat{\pi}\widehat{\delta}_D}\{\mu_{Y0} - \widehat{\Lambda}_Y - \widehat{\delta}(\mu_{D0} - \widehat{\Lambda}_D)\}\}\right]\\
&= P\{q(\widehat{\delta} - \delta_0)\} + P\left[\frac{q}{\widehat{\delta}_D}\frac{(2Z-1)(2T-1)}{\widehat{\pi}}\{\mu_{Y0} - \widehat{\Lambda}_Y - \widehat{\delta}(\mu_{D0} - \widehat{\Lambda}_D)\}\right]\\
&= P\{q(\widehat{\delta} - \delta_0)\} + P\left[\frac{q}{\widehat{\delta}_D}\frac{(2Z-1)(2T-1)}{\widehat{\pi}}\{\Lambda_{Y0} + ZT\delta_0\delta_{D0} - \widehat{\Lambda}_Y - \widehat{\delta}(\Lambda_{D0} + ZT\delta_{D0} - \widehat{\Lambda}_D)\}\right]\\
&= P\{q(\widehat{\delta} - \delta_0)\} + P\left[\frac{q}{\widehat{\delta}_D}\frac{(2Z-1)(2T-1)}{\widehat{\pi}}\{(\Lambda_{Y0} - \widehat{\Lambda}_Y) - \widehat{\delta}(\Lambda_{D0} - \widehat{\Lambda}_D) + ZT\delta_{D0}(\delta_0 - \widehat{\delta})\}\right]\\
&\quad - P\left[\frac{q}{\widehat{\delta}_D}\frac{(2Z-1)(2T-1)}{\pi_0}\{(\Lambda_{Y0} - \widehat{\Lambda}_Y) - \widehat{\delta}(\Lambda_{D0} - \widehat{\Lambda}_D)\}\right]\\
&= P\left[ZTq(\widehat{\delta} - \delta_0)\{\frac{1}{\pi_0} - \frac{\delta_{D0}}{\widehat{\delta}_D\widehat{\pi}}\}\right] + P\left[\frac{q}{\widehat{\delta}_D}\frac{(2Z-1)(2T-1)}{\widehat{\pi}}\{(\Lambda_{Y0} - \widehat{\Lambda}_Y) - \widehat{\delta}(\Lambda_{D0} - \widehat{\Lambda}_D)\}\right]\\
&\quad - P\left[\frac{q}{\widehat{\delta}_D}\frac{(2Z-1)(2T-1)}{\pi_0}\{(\Lambda_{Y0} - \widehat{\Lambda}_Y) - \widehat{\delta}(\Lambda_{D0} - \widehat{\Lambda}_D)\}\right]
\end{aligned}$$

$$
\begin{aligned}
&= P\left[ZTq(\widehat{\delta}-\delta_0)\{\frac{\widehat{\pi}-\pi_0}{\pi_0\widehat{\pi}}+\frac{\widehat{\delta}_D-\delta_{D0}}{\widehat{\delta}_D\widehat{\pi}}\}\right] \\
&\quad + P\left[\frac{q}{\widehat{\delta}_D}(2Z-1)(2T-1)\frac{\pi_0-\widehat{\pi}}{\pi_0\widehat{\pi}}\{(\Lambda_{Y0}-\widehat{\Lambda}_Y)-\widehat{\delta}(\Lambda_{D0}-\widehat{\Lambda}_D)\}\right] \\
&= O_p\Big(\|\widehat{\delta}-\delta_0\|_2(\|\widehat{\pi}-\pi_0\|_2+\|\widehat{\delta}_D-\delta_{D0}\|_2)+\|\widehat{\pi}-\pi_0\|_2(\|\widehat{\Lambda}_Y-\Lambda_{Y0}\|_2+\|\widehat{\Lambda}_D-\Lambda_{D0}\|_2)\Big) \\
&= O_p\Big(\|\widehat{\delta}-\delta_0\|_2(\|\widehat{\pi}-\pi_0\|_2+\|\widehat{\delta}_D-\delta_{D0}\|_2)+\|\widehat{\pi}-\pi_0\|_2(\|\widehat{\Delta}_Y-\Delta_{Y0}\|_2+\|\widehat{\Delta}_D-\Delta_{D0}\|_2)\Big)
\end{aligned}
$$

where the first equality is from (S12) and iterated expectation, the fourth equality is from $E\{(2Z-1)(2T-1)/\pi_0 \mid T, \boldsymbol{X}\} = E\{(2Z-1)(2T-1)/\pi_0 \mid Z, \boldsymbol{X}\} = 0$, the second to the last equality is from the Cauchy-Schwartz inequality that $P(XY) \leqslant \|\boldsymbol{X}\|_2\|Y\|_2$, the boundedness of $q(\boldsymbol{V};\boldsymbol{\psi}_0)$, $1/\widehat{\delta}_D$, $1/\pi_0$, and $1/\widehat{\pi}$ (from the trend relevance assumption, the positivity assumption, and the Donsker condition), and the fact that $(2Z-1)^2(2T-1)^2 = 1$, and the triangle inequality, and the last equality is again from the triangle inequality.

S3.8 *Proof of Theorem S2*

In this section, denote $n_{\min} = \min\{n_a, n_b\}$. From the definition of $\widehat{\beta}_{\mathrm{TS}}$, we have

$$\sqrt{n_{\min}}(\widehat{\beta}_{\mathrm{TS}}-\beta_0) = \frac{\sqrt{n_{\min}}(\widehat{\delta}_{Ya}-\beta_0\widehat{\delta}_{Db})}{\widehat{\delta}_{Db}}.$$

From the two-sample design, $\widehat{\delta}_{Ya}$ is independent of $\widehat{\delta}_{Db}$. Then, similar to the proof of Theorem 2, we can show that

$$\sqrt{n_a}(\widehat{\delta}_{Ya}-\delta_{Ya}) \xrightarrow{d} N\left(0,\ \sum_{t=0}^{1}\sum_{z=0}^{1}\frac{\mathrm{Var}(Y_a \mid T_a=t, Z_a=z)}{P(T_a=t, Z_a=z)}\right),$$

$$\sqrt{n_b}(\widehat{\delta}_{Db}-\delta_{Db}) \xrightarrow{d} N\left(0,\ \sum_{t=0}^{1}\sum_{z=0}^{1}\frac{\mathrm{Var}(D_b \mid T_b=t, Z_b=z)}{P(T_b=t, Z_b=z)}\right).$$

In consequence,

$$
\begin{aligned}
&\sqrt{n_{\min}}\big\{(\widehat{\delta}_{Ya}-\beta_0\widehat{\delta}_{Db})-(\delta_{Ya}-\beta_0\delta_{Db})\big\} \xrightarrow{d} \\
&\qquad N\left(0,\ \sum_{t=0}^{1}\sum_{z=0}^{1}\alpha_a\frac{\mathrm{Var}(Y_a \mid T_a=t, Z_a=z)}{P(T_a=t, Z_a=z)}+\alpha_b\beta_0^2\frac{\mathrm{Var}(D_b \mid T_b=t, Z_b=z)}{P(T_b=t, Z_b=z)}\right).
\end{aligned}
$$

Theorem S2 follows from $\delta_{Ya}-\beta_0\delta_{Db} = \delta_{Ya}-\beta_0\delta_{Da} = 0$, $\widehat{\delta}_{Db} = \delta_{Db}+o_p(1)$ and Slutsky's theorem.

S4. Application

R codes for constructing the dataset and reproducing the results are in smoking-lung.R included in the supplementary materials. In the following, we provide additional details on the application.

S4.1 *Data*

The 1970 NHIS data (*personsx.rds*) were drawn using the R lodown package at `http://asdfree.com`. The CDC mortality data were obtained from the CDC compressed mortality file. The mortality data are also included in the supplementary materials as Compressed Mortality, 1975.txt, Compressed Mortality, 1985.txt, Compressed Mortality, 1995.txt, Compressed Mortality, 2005.txt.

Standard errors for the cigarette smoking prevalence are obtained from the survey package in R to account for the NHIS complex sample design, following the variance estimation procedure available at `https://www.cdc.gov/nchs/data/nhis/6372var.pdf` and also included in the supplementary materials as 6372var.pdf. Standard errors for the lung cancer mortality rates are calculated following `https://wonder.cdc.gov/wonder/help/cmf.html#Standard-Errors`, using the formula $\sqrt{p/n}$, where p is the crude mortality rate, n is the sample size for the population. In Table S1, we include the sample size for each birth cohort in each dataset. According to Theorem S2 and Equation (S2), these obtained standard errors suffice for constructing the consistent variance estimator for $\widehat{\beta}_{\mathrm{TS}}$.

[Table 1 about here.]

S4.2 *Use of gender as a surrogate for encouragement*

It is known that a standard IV does not need to have a causal effect on the exposure (Hernán and Robins, 2006). It is also the case for the IV for DID; the IV for DID Z does not need to

have a causal effect on the exposure; it suffices that the IV for DID is associated with the trend in exposure.

Let D_t be the potential exposure that would be observed at time t if Z takes the value that naturally occurs. Using Z as a surrogate, we can still establish the identification result in Proposition 1 under Assumptions S2 - S3 stated as follows.

ASSUMPTION S2: (a) (Consistency) $D = D_T$ and $Y = Y_T^{(D)}$.

(b) (Positivity) $0 < P(T = t, Z = z|\boldsymbol{X}) < 1$ for $t = 0, 1$, $z = 0, 1$ with probability 1.

(c) (Random sampling) $T \perp (D_t, Y_t^{(d)}, t = 0, 1, d = 0, 1) \mid Z, \boldsymbol{X}$.

ASSUMPTION S3 (Instrumented DID): With probability 1,

(a) (Trend relevance) $\delta_D \neq 0$.

(b) (Independence & exclusion restriction) $Z \perp (Y_1^{(0)} - Y_0^{(0)}, Y_t^{(1)} - Y_t^{(0)}, t = 0, 1)|\boldsymbol{X}$.

(c) (No unmeasured common effect modifier) $E(D_t(Y_t^{(1)} - Y_t^{(0)})|\boldsymbol{X}, Z = 1) - E(D_t(Y_t^{(1)} - Y_t^{(0)})|\boldsymbol{X}, Z = 0) = (E(D_t|\boldsymbol{X}, Z = 1) - E(D_t|\boldsymbol{X}, Z = 0))E(Y_t^{(1)} - Y_t^{(0)})$ for $t = 0, 1$.

(d) (Stable treatment effect over time) $E(Y_1^{(1)} - Y_1^{(0)}|\boldsymbol{X}) = E(Y_0^{(1)} - Y_0^{(0)}|\boldsymbol{X})$.

Note that Assumption S3(c) is implied by Assumption 2. To better understand Assumption S3(c), similar to Wang and Tchetgen Tchetgen (2018), assume in this paragraph only the existence of an unmeasured confounder U_t such that $(D_t, Z) \perp (Y_t^{(1)} - Y_t^{(0)}) \mid U_t, \boldsymbol{X}$ and $Z \perp U_t|\boldsymbol{X}$. Then, the same as the discussion of Theorem 2(c) in the main article, Assumption S3(c) holds if either (i) there is no additive U_t-Z interaction in $E(D_t|Z, U_t, \boldsymbol{X})$: $E(D_t|Z = 1, U_t, \boldsymbol{X}) - E(D_t|Z = 0, U_t, \boldsymbol{X}) = E(D_t|Z = 1, \boldsymbol{X}) - E(D_t|Z = 0, \boldsymbol{X})$; or (ii) there is no additive U_t-d interaction in $E(Y^{(d)}|U_t, \boldsymbol{X})$: $E(Y^{(1)} - Y^{(0)}|U_t, \boldsymbol{X}) = E(Y^{(1)} - Y^{(0)}|\boldsymbol{X})$.

S4.3 *Sensitivity analysis*

As mentioned in the main article, there is still concern about violating the stable treatment effect over time assumption (Assumption 2(d)), possibly because the cigarette design and

composition have undergone changes that promote deeper inhalation of smoke (Thun et al., 2013; Warren et al., 2014). In this section, we apply the sensitivity analysis developed in Section S2.3.

Because the concern is that the effect of smoking on lung cancer increases over time, we consider $\gamma_L = 0$ and $\gamma_U = 0.3\%$, i.e., we consider every value of $\Gamma \in [0, 0.3\%]$. The constructed confidence intervals for each two consecutive birth cohorts are in Figure S1, which indicates that any $\Gamma \in [0, 0.3\%]$ cannot explain away the treatment effect. In fact, any positive Γ cannot explain away the treatment effect. This means that the study conclusion is robust to possible violation of Assumption 2(d).

[Figure 1 about here.]

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies* **72,** 1–19.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* **41,** 1–31.

Fogarty, C. B. (2020). Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association* **115,** 1518–1530.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17,** 360–372.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review Papers and Proceedings* **93,** 126–132.

Neuman, M. D., Rosenbaum, P. R., Ludwig, J. M., Zubizarreta, J. R., and Silber, J. H. (2014). Anesthesia technique, mortality, and length of stay after hip fracture surgery. *JAMA* **311,** 2508–2517.

Richardson, T. S. and Robins, J. M. (2014). Ace bounds; sems with equilibrium conditions. *Statistical Science* **29,** 363–366.

Richardson, T. S., Robins, J. M., and Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association* **112,** 1121–1130.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study. *Statist. Sci.* **2,** 292–306.

Thun, M. J., Carter, B. D., Feskanich, D., Freedman, N. D., Prentice, R., Lopez, A. D., and et al. (2013). 50-year trends in smoking-related mortality in the United States. *New England Journal of Medicine* **368,** 351–364.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.

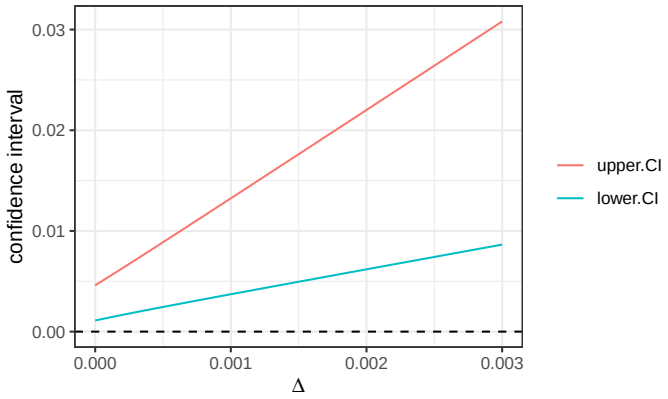
VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine* **167,** 268–274.

Wang, L. and 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.

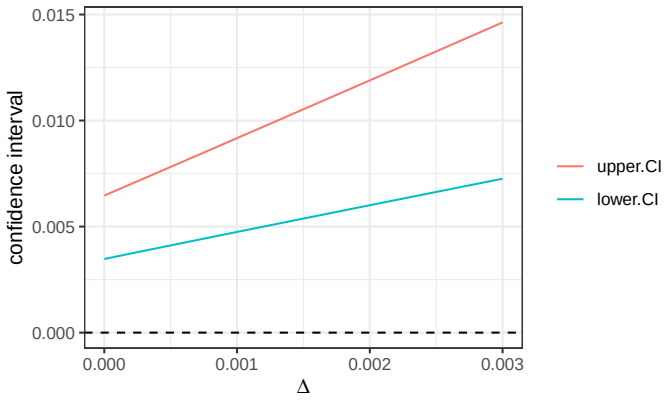
Warren, G. W., Alberg, A. J., Kraft, A. S., and Cummings, K. M. (2014). The 2014 surgeon general's report:"the health consequences of smoking–50 years of progress": a paradigm shift in cancer care. *Cancer* **120,** 1914–1916.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81,** 735–761.

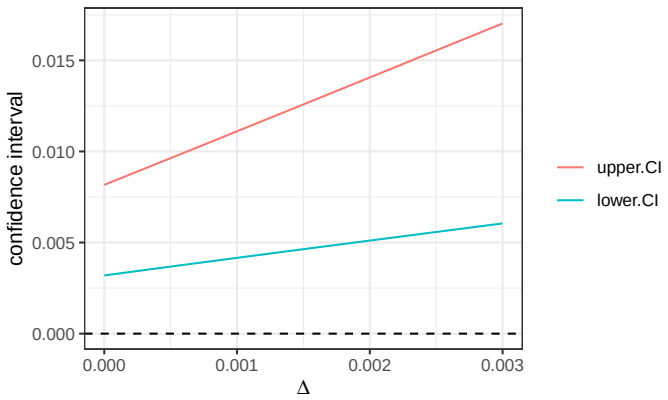
Received April 2022. *Revised April* 2007. *Accepted April* 2007.



(a) Birth cohorts: 1911-1920.



(b) Birth cohorts: 1921-1930.



(c) Birth cohorts: 1931-1940.

Figure S1: Confidence intervals for β^* when $\Gamma \in [0, 0.3\%]$. The confidence intervals do not cover zero, which means that the observed treatment effect cannot be explained away by $\Gamma \in [0, 0.3\%]$.

Table S1: Sample sizes for 1970 NHIS datasets and 1975, 1985, 1995, 2005 CDC WONDER compressed mortality datasets by birth cohort and gender

Birth Cohorts	1911-1920	1921-1930	1931-1940	1941-1950
NHIS				
Men	4,830	5,620	5,343	6,942
Women	6,043	7,024	6,672	8,567
CDC WONDER				
Men	9,416,000	10,383,963	10,158,673	14,773,087
Women	10,629,000	11,751,158	11,161,349	15,868,410