Supporting Information for

Improving efficiency of inference in clinical trials with external control data

by

Xinyu Li, Wang Miao, Fang Lu and Xiao-Hua Zhou

In this Supporting Information, we present all technical proofs of propositions in Web Appendices D–H, and the details of simulation studies for the proposed estimators in Web Appendices K–L. We include detailed descriptions of the study designs and the real datasets of the application in Web Appendices A and B, respectively. We also provide analyses and discussions on the violation of the mean exchangeability in Web Appendix C, the efficiency gain in Web Appendix J and a special case where the trial dataset only contains a treated group in Web Appendix I.

A Discussions on study designs

As discussed in Dahabreh et al. (2020, 2021); Colnet et al. (2020); Li et al. (2021), the study designs for multiple observational datasets generally fall into two categories: (i). nested trial designs (e.g., comprehensive cohort study), in which the trial sample is nested in a large sample that is selected in advance from a well-defined superpopulation; (ii). non-nested trial designs (e.g., composite dataset design), in which observations in different datasets are sampled separately. Our motivating H.pylori application concerns a non-nested one—composite dataset design. As argued by Dahabreh et al. (2020), the nested trial design is generally a census of the underlying population, while the non-nested trial design can be viewed as a biased sampling design.

In this section, we first formally describe the study design of our H.pylori application, and then discuss the relationship to other related study designs, including both nested and non-nested ones. We also clarify the practical implications of the causal quantities of interest when the trial and external dataset are obtained under different study designs.

For a more detailed and thorough description of the study design, analogous to the framework described by Dahabreh et al. (2021); Li et al. (2021), within this section we introduce an additional indicator variable O to indicate whether an individual in the underlying population is sampled and contributes to the observed data $\mathcal{O}$ (trial participants or external controls), with $O = 1$ for sampled individuals and $O = 0$ for non-sampled individuals. We observe information on $(X, Y, T, D = 1, O = 1)$ for trial participants, and information on $(X, Y, T = 0, D = 0, O = 1)$ for external controls. By employing indicator O for selection into the study, we enlarge the observed data $(X, Y, T, D \mid O = 1)$ of sample size n to (OX, OY, OT, OD, O) of sample size N, where N is unknown in non-nested designs. In the following of this section, the observed data are viewed as independent and identically distributed draws from the distribution conditional on $O = 1$; therefore, with some abuse of notation, here we shall write the quantities in the main text as conditional on $O = 1$, e.g., $\pi(X) = \text{pr}(D = 1 \mid X, O = 1)$.

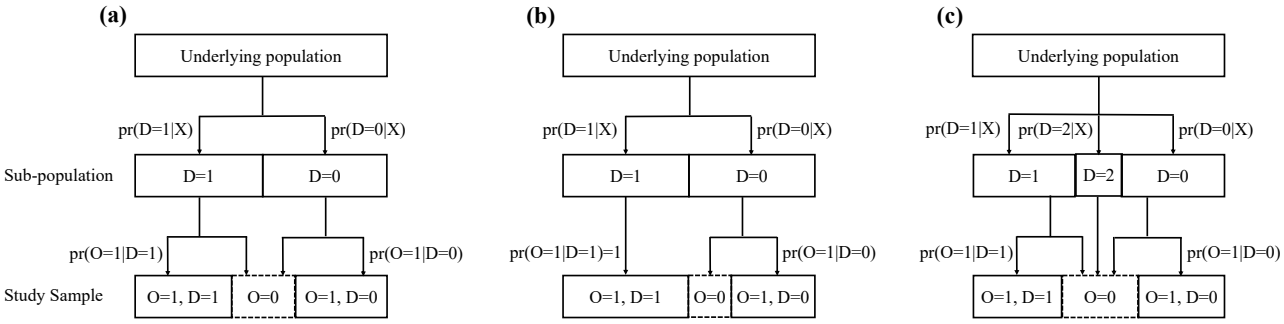

Figure A1: The conceptual frameworks of different study designs: (a) is for our H.pylori application, (b) is for the design adopted by Dahabreh et al. (2021), and (c) is for a more general one that accommodates the design adopted by Dahabreh et al. (2020); Colnet et al. (2020); Li et al. (2021). One can obtain (b) by letting $\text{pr}(O = 1 \mid D = 1) = 1$ in (a), and obtain (a) by letting $\text{pr}(D = 2 \mid X) = 0$ in (c).

We formalize the non-nested study design of our H.pylori application, and illustrate the conceptual relationships between the sampling mechanism and observed study samples in Fig.A1 (a). The underlying population consists of study-eligible individuals to whom research studies would be applicable, also referred to as *actual population* in Dahabreh et al. (2021), with sample size of N from a superpopulation distribution. Two well-defined sub-populations are separated from the underlying population, namely the trial population (or called trial-eligible population, defined as individuals who are able and willing to participate in the *two-arm* trial, $D = 1$) and the external population (or called non-trial eligible population, $D = 0$). The sub-populations are characterized by $\text{pr}(D = 1 \mid X)$ and $\text{pr}(D = 0 \mid X)$, which may be unknown in practice. In our H.pylori infection study, the researchers conducted two separate randomized clinical trials in TCM hospitals and Western-style hospitals, the two major types of hospitals in China. The two-arm clinical trial is limited to only the TCM hospital, where the participants were randomized into two treatments. In the Western-style hospital, the participants were only available to the control treatment—the triple therapy, a standard of care. The two separate clinical trials adopt the same inclusion and exclusion criteria. Therefore, the underlying population is the study-eligible patient population of H.pylori infection who fulfill the inclusion and exclusion criteria and are willing to enroll in the

study, and the two sub-populations are study-eligible patient population at the TCM hospital and that at the Western-style hospital, respectively. In this case, $\text{pr}(D = 1 \mid X)$ and $\text{pr}(D = 0 \mid X)$ can be regarded as capturing preferences of study-eligible patients for hospital type.

Then supposed that we observe a combination of data (size of n) from a subsample of study-eligible patients in the TCM hospital ($D = 1$) and a subsample of study-eligible patients in the Western-style hospital ($D = 0$), with sampling probabilities $\text{pr}(O = 1 \mid D = 1) = u_1 > 0$ and $\text{pr}(O = 1 \mid D = 0) = u_2 > 0$ respectively, where u_1 and u_2 are unknown constants. Note that these sampling conditions imply that $O \perp\!\!\!\perp X, Y, T \mid D$, which is commonly assumed by investigators in non-nested designs, see e.g., Westreich et al. (2017); Dahabreh et al. (2020, 2021). However, our main results, with minor modifications, remain valid when the sampling probabilities are known functions of baseline covariates rather than known constants u_1 and u_2. Allowing the sampling probabilities to depend on baseline covariates, however, does not lead to additional insights regarding study design; for this reason, in the main text, we assume that the sampling probability does not depend on covariates.

The sample size of the observed data n goes to infinite as the sample size of the underlying population $N \to \infty$. We draw inference based on the observed data ($O = 1$), and under the assumptions given in the main text, the causal quantities $\tau = E(Y_1 - Y_0 \mid D = 1, O = 1)$, $\xi = E(Y_1 - Y_0 \mid D = 0, O = 1)$ and $\psi = E(Y_1 - Y_0 \mid O = 1)$ are identifiable. By the design, observations in each dataset are simple random samples from a corresponding sub-population, which suggests that O is independent with other variables given D, and hence we have

$$\begin{aligned} \tau &= E(Y_1 - Y_0 \mid D = 1, O = 1) = E(Y_1 - Y_0 \mid D = 1), \\ \xi &= E(Y_1 - Y_0 \mid D = 0, O = 1) = E(Y_1 - Y_0 \mid D = 0). \end{aligned}$$

This indicates that τ and ξ are the causal quantities that pertain to the study-eligible patient population at the TCM hospital ($D = 1$) and that at the Western-style hospital ($D = 0$), respectively.

Moreover, by the law of total probability, we have

$$\begin{aligned}\psi &= E(Y_1 - Y_0 \mid O = 1) \\ &= E(Y_1 - Y_0 \mid D = 1, O = 1)\text{pr}(D = 1 \mid O = 1) \\ &\quad + E(Y_1 - Y_0 \mid D = 0, O = 1)\text{pr}(D = 0 \mid O = 1) \\ &= E(Y_1 - Y_0 \mid D = 1)\text{pr}(D = 1 \mid O = 1) + E(Y_1 - Y_0 \mid D = 0)\text{pr}(D = 0 \mid O = 1),\end{aligned}$$

where $\text{pr}(D = d \mid O = 1) = \text{pr}(O = 1 \mid D = d)\text{pr}(D = d)/\{\sum_d \text{pr}(O = 1 \mid D = d)\text{pr}(D = d)\}$, $d \in \{0, 1\}$. Therefore, the overall effect ψ is the causal quantity that pertains to a mixed patient population of both the TCM and Western-style hospitals, with mixing probability $\text{pr}(D = 1 \mid O = 1) = \text{pr}(D = 1, O = 1)/\text{pr}(O = 1)$, the limit of n_1/n.

Next we discuss the relationship to other related study designs. Under different study designs, the populations corresponding to $\mathcal{O}_1$ and $\mathcal{O}_2$ may differ, and so may the treatment effects. In the following discussion, we focus on distinguishing the population of each data source in different study designs, rather than the observed data structure; we always assume that information on covariates and outcome is collected in each dataset as described in Table 1 of the paper.

Our study design is quite similar to the non-nested trial design adopted by Dahabreh et al. (2021), which is illustrated in Fig.A1 (b). Under this design, the dataset $\mathcal{O}_1$ is obtained from all trial-eligible individuals, and $\mathcal{O}_2$ from a random sample group of non-trial eligible individuals. The difference lies in that this design requires all trial-eligible individuals to participate in the trial, and accordingly Fig.A1 (b) can be obtained by letting $\text{pr}(O = 1 \mid D = 1) = 1$ in Fig.A1 (a). This holds true for many designs, but may be modified to our study design when, possibly due to restrictions on trial size, researchers have to randomly invite only a part of individuals to enroll in the trial. Because all trial-eligible individuals participate in the trial, the meanings of causal quantities may change slightly: the causal quantity τ pertains to the population of trial participants, ξ pertains to the population of non-participants, and ψ pertains to a mixed population of trial participants and non-participants, see discussion in Dahabreh et al. (2021).

However, as also pointed out by Dahabreh et al. (2021), challenges arise when individuals in the external data are not drawn from the entire non-trial eligible population but from a narrower subset. Dahabreh et al. (2021) discuss a scenario where investigators conduct a randomized trial in the U.S. and also obtain data from members of a private health insurance plan. Due to the existence of other non-overlapping health insurance plans, individuals from the private insurance plan do not represent the entire non-trial eligible population.

To address such problems, we consider a more general conceptual framework of designs. As illustrated in Fig.A1 (c), we separate the underlying population into three sub-populations: trial-eligible individuals ($D = 1$), non-trial eligible individuals from a particular source where data can be available ($D = 0$), and other non-trial eligible individuals outside the range of the first two data sources ($D = 2$). By taking into account the sub-population where data are not available, we relax the restriction $\text{pr}(D = 1 \mid X) + \text{pr}(D = 0 \mid X) = 1$ and thereby avoid the problems discussed above. Pearl et al. (2014) also consider the case of multiple sub-populations. Fig.A1 (a) can be viewed as a special case of Fig.A1 (c) obtained by letting $\text{pr}(D = 2 \mid X) = 0$.

The non-nested design considered by Dahabreh et al. (2020); Colnet et al. (2020); Li et al. (2021) can be accommodated in the framework described by Fig.A1 (c). In their design, the individuals of external data $\mathcal{O}_2$ are simple random samples from the underlying population, which is referred to as the *target population* since investigators are interested in drawing inference on it. Such a design can be obtained by letting $\text{pr}(D = 0 \mid X)$ be a constant in Fig.A1 (c). We then can identify the distribution of baseline covariates in underlying population from the external data. In this design, τ is the causal quantity of interest that pertains to the population of trial participants, and ξ is the one that pertains to the underlying population (target population), but ψ is commonly not of interest to researchers.

At last, we discuss the relationships to nested designs, in which the trial sample (size of n) is selected from and nested in the underlying population. As illustrated in Dahabreh et al. (2021), nested designs can be characterized by Fig.A1 (a), with $\text{pr}(O = 1 \mid D = d)$ being known constants for $d \in \{0, 1\}$. For example, in the comprehensive cohort study design introduced by Olschewski

and Scheurlen (1985), individuals from a well-defined underlying population are first invited to participate in a randomized trial, after which those who refuse would be asked to join a parallel observational study that allows them to choose treatment by their own preferences. Note $\text{pr}(O = 1 \mid D = d) = 1$ for comprehensive cohort studies, and we have $N = n$. Therefore the *overall population* corresponding to all individuals of the study is exactly the underlying population, which generally does not hold in non-nested designs. In comprehensive cohort studies, the causal quantities of interest are τ and ψ, which pertain to the population of nested trial participants and the underlying population, respectively; whereas ξ is not of interest to researchers.

Table A1: A summary of different study designs for multiple observational datasets

Design	Sampling mechanism	Populations to which causal quantities correspond	Observed data properties
non-nested design of the H.pylori application	Fig.A1 (a)	τ trial-eligible population ξ non-trial eligible population ψ a mixed population of both	a random sample of the trial-eligible population combined with a random sample of the non-trial eligible population
non-nested design considered by Dahabreh et al. (2021)	Fig.A1 (b)	τ trial participants population ξ non-participants population ψ a mixed population of both	trial participants combined with a simple random sample of the non-participants
non-nested design considered by Li et al. (2021)	Fig.A1 (c) with $\text{pr}(D = 0 \mid X) > 0$ being a constant	ξ underlying population (target population)	individuals of the external data represent the target population
comprehensive cohort study design (nested)	Fig.A1 (b) with $\text{pr}(O = 1 \mid D = 0) = 1$	τ trial participants population ψ underlying population	all individuals contributing data to the analysis are a census of the underlying population

B Descriptions of the real data

In the motivating H.pylori infection application, the trial dataset is obtained from a two-arm clinical trial conducted at the TCM hospital, where the participants were randomized into two treatments. The external dataset is obtained from a single-arm clinical trial conducted at the Western-style hospital located in Beijing, where participants were only available to the control treatment. In the study, the control treatment is the triple therapy (clarithromycin, amoxicillin and omeprazole), which is a standard of care widely used for H.pylori infection regardless of the type of hospital and participation the study. The active treatment is a combination treatment including both the triple therapy and the TCM herbal therapy (phytotherapy). The trial dataset contains 362 samples, of which 180 patients are assigned to the triple therapy, and the rest are assigned to the combination treatment; the external dataset contains 110 control samples.

The two separate randomized clinical trials adopted the same inclusive and exclusive criteria, and shared the same treatment protocols. For at least 2 weeks prior to treatment, patients were instructed not to take any other medications that may affect the clinical result or interact with a study drug, such as PPI, H2RA and bismuth. During the study, researchers examined the medical records and recorded the patients' general information including age, gender, height and body mass index (BMI), as well as other relevant factors at the societal level such as work type, education level and marriage status. Besides, researchers collected information on health condition before and after treatment by scoring the Patient-Reported Outcome(PRO) measure, and evaluated patients' symptoms that are related to H.pylori infection (e.g., degree of stomach ache, heartburn, acid reflux, burping, nausea and vomiting). The baseline covariates are the same in both external and trial datasets. The treatment process lasted for four weeks, and the outcome of interest is the binary disease status after the treatment detected by the C-14 urea breath test (UBT).

The variable definitions of outcome and pre-treatment baseline covariates are as follows. We also provide a list of descriptive statistics of the trial and external dataset in Table B1.

Name	**Label**
ID	*sequential id runs from 1 to 472*
Treatemt	*1 if treated group*
Data.source	*1 if two-arm randomized clinical trial*
Outcome	*1 if tested for no H.pylori bacterial infection after therapy*
Age	*age*
Pro.value	*pre-treatment Patient-Reported Outcome(PRO) measure on health condition*
Height	*height*
BMI	*body mass index*
Gender	*1 if female*
Education	*education level 1-4*
Marriage	*marriage status*
Ethnicity	*1 if the non-Han ethnic group*
Job	*work type*
Gastritis	*gastritis type*
After.care	*1 if taking medicines prior to the admission to hospital*
Stomachache	*degree of stomach ache 1-3*
Heartburn	*1 if heartburn or acid reflux*
Burping	*1 if burping*
Nausea	*1 if nausea or vomiting*

Table B1: Descriptive statistics of pre-treatment baseline covariates

Age	con.					
		$\mathcal{O}_1$	Min: 18	1st Qu: 33	Median: 45	3rd Qu: 54	Max: 64
		$\mathcal{O}_2$	Min: 20	1st Qu: 34	Median: 48	3rd Qu: 56	Max: 65

Pro.value	con.					
		$\mathcal{O}_1$	Min: 40	1st Qu: 65	Median: 74	3rd Qu: 88	Max: 150
		$\mathcal{O}_2$	Min: 47	1st Qu: 66	Median: 77.5	3rd Qu: 88	Max: 131
Height	con.					
		$\mathcal{O}_1$	Min: 148	1st Qu: 160	Median: 165	3rd Qu: 170	Max: 185
		$\mathcal{O}_2$	Min: 146	1st Qu: 160	Median: 164	3rd Qu: 172	Max: 181
BMI	con.					
		$\mathcal{O}_1$	Min: 16.1	1st Qu: 20.5	Median: 22.5	3rd Qu: 24.5	Max: 34.1
		$\mathcal{O}_2$	Min: 17.2	1st Qu: 20.8	Median: 22.9	3rd Qu: 25.0	Max: 35.3
Gender	catg.					
		$\mathcal{O}_1$	0 : 148	1 : 214		
		$\mathcal{O}_2$	0 : 42	1 : 68		
Education	catg.					
		$\mathcal{O}_1$	1 : 32	2 : 68	3 : 83	4 : 179
		$\mathcal{O}_2$	1 : 3	2 : 14	3 : 23	4 : 70

Marriage	catg.			
		$\mathcal{O}_1$	0 : 47	1 : 315
		$\mathcal{O}_2$	0 : 17	1 : 93
Ethnicity	catg.			
		$\mathcal{O}_1$	0 : 352	1 : 10
		$\mathcal{O}_2$	0 : 105	1 : 5
Job	catg.			
		$\mathcal{O}_1$	0 : 256	1 : 97
		$\mathcal{O}_2$	0 : 88	1 : 22
Gastritis	catg.			
		$\mathcal{O}_1$	0 : 263	1 : 99
		$\mathcal{O}_2$	0 : 91	1 : 19
After.care	catg.			
		$\mathcal{O}_1$	0 : 277	1 : 85
		$\mathcal{O}_2$	0 : 85	1 : 25

Stomachache	catg.			
	$\mathcal{O}_1$	1 : 79	2 : 146	3 : 137
	$\mathcal{O}_2$	1 : 48	2 : 37	3 : 25

Heartburn	catg.		
	$\mathcal{O}_1$	0 : 132	1 : 230
	$\mathcal{O}_2$	0 : 53	1 : 57

Burping	catg.		
	$\mathcal{O}_1$	0 : 98	1 : 264
	$\mathcal{O}_2$	0 : 43	1 : 67

Nausea	catg.		
	$\mathcal{O}_1$	0 : 279	1 : 83
	$\mathcal{O}_2$	0 : 91	1 : 19

C Discussions on the violation of mean exchangeability

The mean exchangeability plays a key role for integration of information across the trial and external datasets. Although it is testable, violation of the mean exchangeability assumption can lead to bias when the external data are used to improve efficiency of inference. In this sense, incorporation of external data can be viewed as a trade-off between bias and variance—potential bias may arise along

with efficiency gains. Therefore, it is essential to analyze the impact of potential bias introduced by external data, assess the sensitivity to violation of mean exchangeability, and discuss this issue more in-depth.

To better understand the variance and bias trade-off due to the incorporation of external control data, we provide an analysis of the potential bias, propose a test of the mean exchangeability, and conduct additional numerical simulations particularly under the scenarios where the mean exchangeability is violated. Specifically, we demonstrate how the violation of mean exchangeability will impact the asymptotic bias of the full-data estimator $\hat{\tau}_{\text{dr}}$, and we show that in certain cases where the engagement effects are weak, this estimator may still have a smaller MSE than the trial-based doubly robust estimator. In the presence of strong engagement effects, the bias is no longer negligible, however, can be detected via the proposed test of the mean exchangeability. In our H.pylori application, the test of the mean exchangeability yields a p-value of 0.441. Hence under the significance level of 0.05, we do not reject the mean exchangeability.

C.1 Bias analysis

We denote the selection bias by

$$E(Y_0 \mid X, D = 1) - E(Y_0 \mid X, D = 0) = b(X),$$

which encodes the strength of engagement effects (effects of participation in a particular study on the outcome that are not through treatment) within each level of X. Assuming sufficiently flexible working models are employed such that no approximation error is introduced by model misspecification, we can show that the asymptotic bias of the full-data estimator $\hat{\tau}_{\text{dr}}$ is

$$\Lambda = E\left[\frac{\pi(X)}{\text{pr}(D=1)} \cdot \frac{\{1-\pi(X)\}r(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)} \cdot b(X)\right]. \tag{S.1}$$

This equation demonstrates the impact of violation of the mean exchangeability on the asymptotic bias of the full-data estimator $\hat{\tau}_{\text{dr}}$. Note that Λ can occasionally be zero even if $b(X)$ is not zero.

Suppose $b(X)$ is bounded with $|b(X)| \leq B$, then from (S.1), we have $|\Lambda| \leqslant B$, which states that the asymptotic bias of $\hat{\tau}_{\text{dr}}$ does not exceed that the largest difference between the conditional means of two datasets. As a result, weak engagement effects would not negate our inference, although large ones can.

C.2 Simulations under settings where mean exchangeability fails

We evaluate the performance of the proposed method with numerical simulations under settings where mean exchangeability fails. We first consider the data generating mechanism described in the scenario (i) of the simulation in the main text. The outcome variable Y is continuous. We set $b(X) = \delta E(Y_0 \mid X, D = 1)$, and consider the cases where δ takes values in $\pm(0\%, 5\%, 10\%, 20\%)$. We simulate 1000 replicates under 500 sample size for each case and summarize the results with boxplots in Fig.C1. Fig.C2 shows the mean-squared errors (MSE) of the proposed estimator $\hat{\tau}_{\text{dr}}$ at different levels of bias. The MSE grows as the size of bias increases; yet when the size of bias is smaller than 10% of the conditional mean, the full-data doubly robust estimator is still advantageous in terms of MSE.

We also conduct simulations for the binary outcome case. We generate the baseline covariates $X = (X_1, X_2)$ by a multivariate normal distribution with mean $(0, 0)$, variance $(1, 1)$ and correlation $\rho = 0.1$, and generate (D, T, Y) by the following mechanism:

$$\begin{cases} D \mid X \sim \text{Ber}([1 + \exp\{-(-0.5 - 0.5X_1 - 0.2X_2)\}]^{-1}), \\ T \mid (X, D = 1) \sim \text{Ber}([1 + \exp\{-(1 - X_1 + 0.3X_2)\}]^{-1}); \end{cases}$$

$$\begin{cases} Y \mid (X, D = 1, T = 1) \sim \text{Ber}([1 + \exp\{-(2 + 0.5X_1 + X_2)\}]^{-1}), \\ Y \mid (X, D = 1, T = 0) \sim \text{Ber}([1 + \exp\{-(-0.5X_1 - X_2)\}]^{-1}), \\ Y \mid (X, D = 0, T = 0) \sim \text{Ber}([1 + \exp\{-(\delta - 0.5X_1 - X_2)\}]^{-1}); \end{cases}$$

where δ takes values in $\pm(0, 0.1, 0.2, 0.5)$. The value of δ characterizes the difference between conditional outcome means in the log odds ratio scale. Fig.C1 summarizes the results for the bias of estimators. Fig.C2 shows the MSE of the proposed estimator $\hat{\tau}_{\text{dr}}$ at different levels of bias. For

$\delta = \pm 0.5$, the full-data estimator performs analogously to the trial-based estimator in terms of MSE, but has a smaller MSE when $|\delta| < 0.5$.

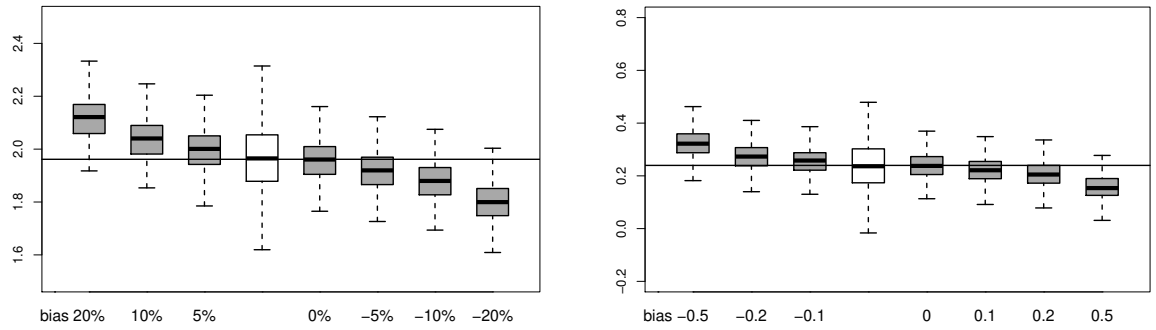

Figure C1: *Boxplots for estimators of the treatment effect τ at different levels of bias. The left is for a continuous outcome, and the right is for a binary outcome. The white box is for the trial-based doubly robust estimator $\tilde{\tau}_{\mathrm{dr}}$, and the grey ones are for the full-data doubly robust estimator $\hat{\tau}_{\mathrm{dr}}$. The horizontal line labels the true value.*

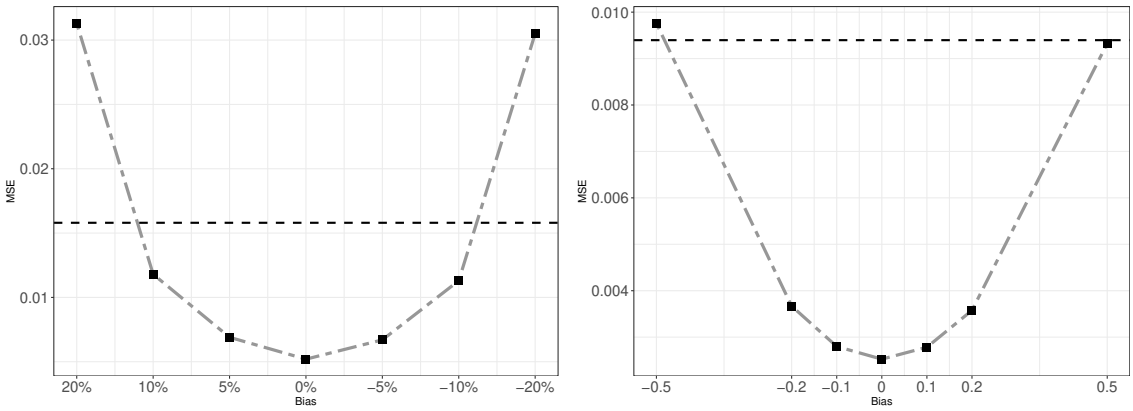

Figure C2: *MSE of the estimator $\hat{\tau}_{\mathrm{dr}}$ at different levels of bias. The left is for a continuous outcome, and the right is for a binary outcome. The horizontal dashed line labels the MSE of the trial-based doubly robust estimator $\tilde{\tau}_{\mathrm{dr}}$.*

C.3 Testing of the mean exchangeability

Here we focus on testing of the mean exchangeability while assuming strong ignorability in the clinical trial, which offers bias detection and robustness assessment of the full-data estimators.

Under ignorability and given the data structure, we have $E(Y_0 \mid D = 1, X) = E(Y \mid D = 1, T = 0, X)$ and $E(Y_0 \mid D = 0, X) = E(Y \mid D = 0, T = 0, X)$. Therefore, a straightforward test

of the mean exchangeability is to test whether $E(Y \mid D=1, T=0, X) = E(Y \mid D=0, T=0, X)$ holds. There has been a few researches on testing the equality between functions, e.g., see Luedtke et al. (2019) for a nonparametric omnibus approach. For a parametric approach, we employ the following model,

$$E(Y \mid X, D, T=0) = g(\beta_0 + \beta_1 D + \beta_2^{\mathrm{T}} X + \beta_3^{\mathrm{T}} DX),$$

where $g(\cdot)$ is a known link function, such as the identity function for a continuous outcome and the logit link for a binary one. The coefficients $(\beta_1, \beta_3^{\mathrm{T}})$ encode the departure from the mean exchangeability. Under this model, to investigate whether the mean exchangeability holds, we test $H_0 : (\beta_1, \beta_3^{\mathrm{T}}) = 0$ versus the alternative $H_1 : (\beta_1, \beta_3^{\mathrm{T}}) \neq 0$, i.e. at least a dimension of the interaction coefficient is not equal to zero. Note that such a test relies on correct specification of the working model.

An alternative test directly addresses the potential bias of the full-data estimator is the Hausman type test, where the test statistic $h = \frac{(\tilde{\tau}_{\mathrm{dr}} - \hat{\tau}_{\mathrm{dr}})^2}{\widehat{\mathrm{var}}(\tilde{\tau}_{\mathrm{dr}}) - \widehat{\mathrm{var}}(\hat{\tau}_{\mathrm{dr}})}$ is asymptotically distributed as $\chi^2(1)$ under the null hypothesis that both estimators are consistent with correct working models. A large value of h indicates a large deviation between these two estimators and is evidence of bias and violation of the mean exchangeability.

We illustrate the tests with simulations with a continuous outcome under the setting described in C.2. At different level of bias, we simulate 1000 replicates under 500, 1000 and 2000 sample size, respectively. The significance level to reject the null hypothesis is set a priori at 0.05 in all cases. Fig.C3 summarizes the power of the tests. When the mean exchangeability holds, the type I error is approximately the nominal significance level of 0.05. When the mean exchangeability fails, the power of the test increases with the size of bias and approximates to one as the sample size increases.

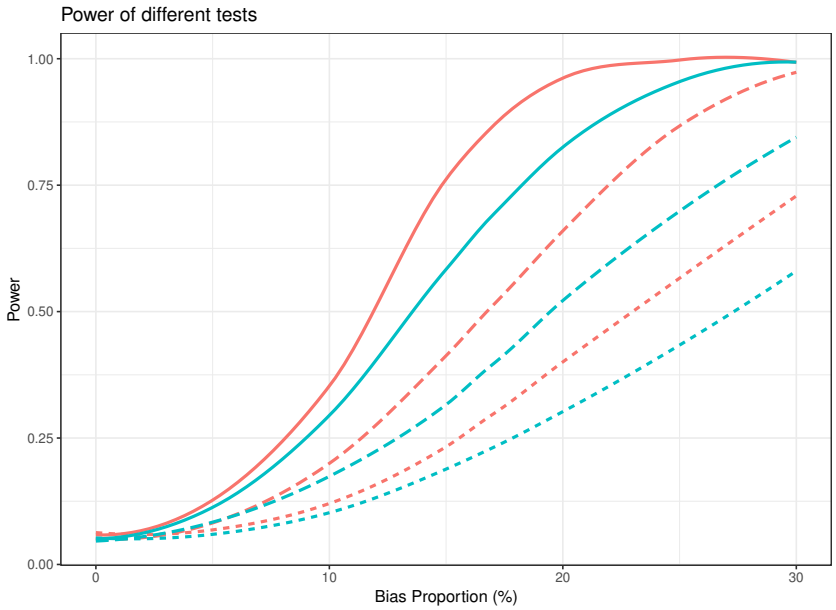

Figure C3: Power of the two proposed tests at different levels of bias under sample size 500, 1000 and 2000. The red and blues curves are for the parametric test and the Hausman type test, respectively. The dashed, longdashed, and solid curves are for sample size 500, 1000, and 2000, respectively.

C.4 Summary

We summarize the points concerning the violation of mean exchangeability: The full-data estimator $\hat{\tau}_{\text{dr}}$ generally fails to be consistent when mean exchangeability is violated. However, its asymptotic bias would not exceed the maximum absolute value of selection bias, which enables us to assess robustness of the inference via sensitivity analysis. In situations where the engagement effects are weak, the full-data doubly robust estimator may enjoy a smaller MSE than the trial-based doubly robust estimator. For detecting the potential bias due to violation of mean exchangeability, we can test whether the assumption holds by testing $E(Y \mid X, D = 1, T = 0) = E(Y \mid X, D = 0, T = 0)$ based on the observed data, or use the Hausman type test. The tests perform well in the simulation studies; when the mean exchangeability fails, the power of tests increases to one as the sample size increases.

At last, we provide a proof of (S.1).

Proof of (S.1). For ease of reference, we let $m_d(X) = E(Y_d \mid X, D = 1)$ for $d \in \{0, 1\}$. Under the mean exchangeability assumption, we have $m_0(X) = E(Y \mid X, T = 0)$, and thereby we can fit the outcome model $m_0(X; \beta_0)$ based on control samples from both the trial and external data. We denote the probability limit of resulting estimator $m_0(X; \hat{\beta}_0)$ by $m_0^*(X)$. However, when the mean exchangeability fails, $m_0^*(X)$ is not necessarily equal to $m_0(X)$.

Now we consider the case where $b(X) \neq 0$. Under standard regularity conditions for estimating equations, the full-data doubly robust estimator $\hat{\tau}_{\text{dr}}$ would converge in probability to

$$\begin{aligned}\tau^* &= \frac{1}{\text{pr}(D=1)} E\Big[D\{m_1(X) - m_0^*(X)\} + \frac{DT}{p(X)}\{Y - m_1(X)\} \\ &\qquad - \pi(X)\frac{D(1-T) + (1-D)r(X)}{\pi(X)\{1-p(X)+\{1-\pi(X)\}r(X)}\{Y - m_0^*(X)\}\Big].\end{aligned}$$

Therefore we can derive the asymptotic bias as

$$\begin{aligned}\Lambda &= \tau^* - \tau \\ &= \tau^* - \frac{1}{\text{pr}(D=1)} E\Big[D\{m_1(X) - m_0(X)\}\Big] \\ &= \frac{1}{\text{pr}(D=1)} E\Big[D\{m_0(X) - m_0^*(X)\} \\ &\qquad - \pi(X)\frac{D(1-T) + (1-D)r(X)}{\pi(X)\{1-p(X)+\{1-\pi(X)\}r(X)}\{Y - m_0^*(X)\}\Big] \\ &= \frac{1}{\text{pr}(D=1)} E\Big[D\{m_0(X) - m_0^*(X)\} \\ &\qquad -\pi(X)\frac{D(1-T)}{\pi(X)\{1-p(X)+\{1-\pi(X)\}r(X)}\{m_0(X) - m_0^*(X)\} \\ &\qquad - \pi(X)\frac{(1-D)r(X)}{\pi(X)\{1-p(X)+\{1-\pi(X)\}r(X)}\{m_0(X) - b(X) - m_0^*(X)\}\Big] \\ &= \frac{1}{\text{pr}(D=1)} E\left[\pi(X)\cdot\frac{\{1-\pi(X)\}r(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\cdot b(X)\right].\end{aligned}$$

□

D Proof of proposition 1

Denote the density of X by $f(x)$, the conditional density $f(Y_j = y \mid X = x, D = 1)$ by $f_j(y \mid x)$ for $j = 0$ or 1, and $f(Y_0 = y \mid X = x, D = 0)$ by $f_0^*(y \mid x)$.

The full observational data distribution is

$$f(y, x, d, t) = [f_1(y \mid x)p(x)]^{dt}[f_0(y \mid x)\{1 - p(x)\}]^{d(1-t)}[f_0^*(y \mid x)]^{1-d}\pi(x)^d[1 - \pi(x)]^{1-d}f(x).$$

Under assumptions 1–3, we consider a regular parametric submodel indexed by θ as

$$[f_1(y \mid x, \theta)p(x, \theta)]^{dt}[f_0(y \mid x, \theta)\{1 - p(x, \theta)\}]^{d(1-t)}[f_0^*(y \mid x, \theta)]^{1-d}\pi(x, \theta)^d[1 - \pi(x, \theta)]^{1-d}f(x, \theta),$$

which equals $f(y, x, d, t)$ when $\theta = \theta_0$ and satisfies $\int yf_0(y \mid x, \theta)dy = \int yf_0^*(y \mid x, \theta)dy$ for $\forall x$.

Then we derive the corresponding score $S(y, x, d, t \mid \theta)$ as

$$\begin{aligned} S(y, x, d, t \mid \theta) &= dtS_1(y \mid x, \theta) + d(1 - t)S_0(y \mid x, \theta) + (1 - d)S_0^*(y \mid x, \theta) \\ &\quad + \frac{d - \pi(x, \theta)}{\pi(x, \theta)\{1 - \pi(x, \theta)\}}\dot{\pi}(x, \theta) + \frac{dt - p(x, \theta)d}{p(x, \theta)\{1 - p(x, \theta)\}}\dot{p}(x, \theta) + S_f(x, \theta), \end{aligned}$$

where $S_j(y \mid x, \theta) = \partial \log f_j(y \mid x, \theta)/\partial\theta$ for $j = 0$ or 1, $S_0^*(y \mid x, \theta) = \partial \log f_0^*(y \mid x, \theta)/\partial\theta$, $S_f(x, \theta) = \partial \log f(x, \theta)/\partial\theta$, and $\dot{\pi}(x, \theta)$, $\dot{p}(x, \theta)$ are pathwise derivatives for $\pi(x, \theta)$ and $p(x, \theta)$.

By the score, the tangent space is

$$\mathcal{T} = \left\{ dtS_1(y \mid x) + d(1 - t)S_0(y \mid x) + (1 - d)S_0^*(y \mid x) + [d - \pi(x)]a(x) + d[t - p(x)]b(x) + S(x) \right\},$$

where $S_j(y \mid x)$ satisfies $\int S_j(y \mid x)f_j(y \mid x)dy = 0$ for $\forall x, j = 0$ or 1, $S_0^*(y \mid x)$ satisfies $\int S_0^*(y \mid x)f_0^*(y \mid x)dy = 0$ and $\int yS_0(y \mid x)f_0(y \mid x)dy = \int yS_0^*(y \mid x)f_0^*(y \mid x)dy$ for $\forall x$, in addition $S(x)$ satisfies $\int S(x)f(x)dx = 0$, and $a(x), b(x)$ are arbitrary square-integrable measurable functions.

Under the parametric submodel indexed by θ, the average treatment effect among the population corresponding to the trial data can be written in the form as

$$\tau(\theta) = \frac{\iint y f_1(y \mid x, \theta)\pi(x, \theta)f(x, \theta)dydx - \iint y f_0(y \mid x, \theta)\pi(x, \theta)f(x, \theta)dydx}{\int \pi(x, \theta)f(x, \theta)dx},$$

where $\tau = \tau(\theta_0)$. For simplicity, we denote $\dot{\pi}(x) = \dot{\pi}(x, \theta_0)$, $\dot{p}(x) = \dot{p}(x, \theta_0)$, $S_f(x) = S_f(x, \theta_0)$, $q = \int \pi(x)f(x)dx$ and $\Delta(x) = \int y f_1(y \mid x)dy - \int y f_0(y \mid x)dy$. Then we have

$$\begin{aligned}\left.\frac{\partial\tau(\theta)}{\partial\theta}\right|_{\theta=\theta_0} &= \frac{\iint y\pi(x)S_1(y \mid x, \theta_0)f_1(y \mid x)f(x)dydx - \iint y\pi(x)S_0(y \mid x, \theta_0)f_0(y \mid x)f(x)dydx}{q} \\ &\quad + \frac{\int\{\Delta(x) - \tau\}\dot{\pi}(x)f(x)dx}{q} + \frac{\int\{\Delta(x) - \tau\}\pi(x)S_f(x)f(x)dx}{q}.\end{aligned}$$

We let

$$\begin{aligned}\mathrm{IF}_\tau(Y, X, D, T) &= \frac{1}{q}\left[D\Delta(X) - D\tau + \frac{DT}{p(X)}\{Y - m_1(X)\}\right. \\ &\quad - \frac{D(1-T)\pi(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\{Y - m_0(X)\} \\ &\quad \left. -(1-D)\frac{\pi(X)r(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\{Y - m_0(X)\}\right].\end{aligned}$$

By tedious calculations given below, pathwise differentiability is satisfied as we can verify that

$$\left.\frac{\partial\tau(\theta)}{\partial\theta}\right|_{\theta=\theta_0} = E\Big\{\mathrm{IF}_\tau(Y, X, D, T) \cdot S(Y, X, D, T \mid \theta_0)\Big\}. \tag{S.2}$$

In addition to (S.2), we can also verify that $\mathrm{IF}_\tau(Y, X, D, T)$ belongs to the tangent space $\mathcal{T}$, which indicates that it is the efficient influence function for the parameter τ in the nonparametric model, and the semiparametric efficiency bound for regular and asymptotic linear estimators of the parameter τ is the expected square of $\mathrm{IF}_\tau(Y, X, D, T)$, that is $B_\tau = E\{\mathrm{IF}_\tau(Y, X, D, T)^2\}$.

To complete the proof, now we provide a verification for the equality (S.2) in the following.

By assumptions 1–3 and lemma 2, $E\{Y - m_1(X) \mid D = 1, T = 1, X\} = E\{Y - m_0(X) \mid D = 1, T = 0, X\} = E\{Y - m_0(X) \mid D = 0, X\} = 0$, then we have

$$
\begin{aligned}
& E\left\{ \mathrm{IF}_\tau(Y, X, D, T) \cdot S(Y, X, D, T \mid \theta_0) \right\} \\
= \; & \frac{1}{q} E\Big[D\{\Delta(X) - \tau\} \cdot S(Y, X, D, T \mid \theta_0) && (E.1) \\
& + DT \left\{ \frac{Y - m_1(X)}{p(X)} \cdot S_1(Y \mid X, \theta_0) \right\} && (E.2) \\
& - D(1-T) \left\{ \frac{Y - m_0(X)}{\pi(X)\{1 - p(X)\} + \{1 - \pi(X)\} r(X)} \pi(X) \cdot S_0(Y \mid X, \theta_0) \right\} && (E.3) \\
& - (1-D) \left\{ \frac{Y - m_0(X)}{\pi(X)\{1 - p(X)\} + \{1 - \pi(x)\} r(X)} \pi(X) r(X) \cdot S_0^*(Y \mid X, \theta_0) \right\} \Big], && (E.4)
\end{aligned}
$$

where

$$
\begin{aligned}
(E.1) \quad & \overset{*}{=} \quad E\Big[D\{\Delta(X) - \tau\} \cdot \frac{D - \pi(X)}{\pi(X)\{1 - \pi(X)\}} \dot{\pi}(X) \\
& \qquad + D\{\Delta(X) - \tau\} \cdot \frac{DT - p(X)D}{p(X)\{1 - p(X)\}} \dot{p}(X) \\
& \qquad + D\{\Delta(X) - \tau\} \cdot S_f(X) \Big] \\
& = \quad E\Big[\{\Delta(X) - \tau\} \dot{\pi}(X) + 0 + \{\Delta(X) - \tau\} \pi(X) S_f(X) \Big], \\
(E.2) \quad & = \quad E\Big[\pi(X) E\Big\{ (Y - m_1(X)) \cdot S_1(Y \mid X, \theta_0) \mid D = 1, T = 1, X \Big\} \Big] \\
& \overset{**}{=} \quad E\Big[\pi(X) E\Big\{ Y \cdot S_1(Y \mid X, \theta_0) \mid D = 1, T = 1, X \Big\} \Big], \\
(E.3) + (E.4) \quad & \overset{**}{=} \quad -E\left[\frac{\pi^2(X)\{1 - p(X)\}}{\pi(X)\{1 - p(X)\} + \{1 - \pi(X)\} r(X)} E\Big\{ Y \cdot S_0(Y \mid X, \theta_0) \mid D = 1, T = 0, X \Big\} \right] \\
& \qquad -E\left[\frac{\pi(X)\{1 - \pi(X)\} r(X)}{\pi(X)\{1 - p(X)\} + \{1 - \pi(X)\} r(X)} E\Big\{ Y \cdot S_0^*(Y \mid X, \theta_0) \mid D = 0, X \Big\} \right] \\
& \overset{***}{=} \quad -E\Big[\pi(X) E\Big\{ Y \cdot S_0(Y \mid X, \theta_0) \mid D = 1, T = 0, X \Big\} \Big].
\end{aligned}
$$

Therefore, we obtain that

$$\begin{aligned} & E\Big\{\mathrm{IF}_\tau(Y,X,D,T)\cdot S(Y,X,D,T\mid\theta_0)\Big\} \\ = \;& \frac{1}{q}\Big\{(E.1)+(E.2)+(E.3)+(E.4)\Big\} \\ = \;& \frac{\partial\tau(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}. \end{aligned}$$

* the other terms equal zero because $\int S_j(y\mid x,\theta_0)f_j(y\mid x)dy=0$ for $j=0$ or 1.

** the conditional means can be omitted because $\int S_j(y\mid x,\theta_0)f_j(y\mid x)dy=0$ for $j=0$ or 1, and $\int S_0^*(y\mid x,\theta_0)f_0^*(y\mid x)dy=0$.

*** the equality holds because $\int yS_0(y\mid x,\theta_0)f_0(y\mid x)dy=\int yS_0^*(y\mid x,\theta_0)f_0^*(y\mid x)dy$.

In addition, we verify that $\mathrm{IF}_\tau(Y,X,D,T)$ belongs to the tangent space $\mathcal{T}$.

Let

$$\begin{aligned} S_1(y\mid x) &= \frac{1}{q}\cdot\frac{1}{p(x)}\left\{y-m_1(x)\right\}, \\ S_0(y\mid x) &= -\frac{1}{q}\cdot\frac{\pi(x)}{\pi(x)\{1-p(x)\}+\{1-\pi(x)\}r(x)}\{Y-m_0(x)\}, \\ S_0^*(y\mid x) &= -\frac{1}{q}\cdot\frac{\pi(x)r(x)}{\pi(x)\{1-p(x)\}+\{1-\pi(x)\}r(x)}\{Y-m_0(x)\}, \\ a(x) &= \frac{1}{q}\cdot\{\Delta(x)-\tau\}, \\ b(x) &= 0, \\ S(x) &= \frac{1}{q}\cdot\pi(x)\{\Delta(x)-\tau\}, \end{aligned}$$

then

$$\mathrm{IF}_\tau(y,x,d,t)=dtS_1(y\mid x)+d(1-t)S_0(y\mid x)+(1-d)S_0^*(y\mid x)+\{d-\pi(x)\}a(x)+d\{t-p(x)\}b(x)+S(x),$$

and the functions satisfy

$$\begin{aligned}
\int S_1(y \mid x) f_1(y \mid x) dy &= 0, \\
\int S_0(y \mid x) f_0(y \mid x) dy &= 0, \\
\int S_0^*(y \mid x) f_0^*(y \mid x) dy &= 0, \\
\int y S_0(y \mid x) f_0(y \mid x) dy &= \int y S_0^*(y \mid x) f_0^*(y \mid x) dy, \\
\int S(x) f(x) dx &= 0,
\end{aligned}$$

which completes the proof.

E Proof of the equality (1)

The trial-based doubly robust estimator is

$$\tilde{\tau}_{\mathrm{dr}} = \hat{E}\left[\left\{ m_1(X;\tilde{\beta}_1) - m_0(X;\tilde{\beta}_0) + \frac{T}{p(X;\tilde{\phi})}\tilde{R}_1 - \frac{1-T}{1-p(X;\tilde{\phi})}\tilde{R}_0 \right\} \mid D = 1\right],$$

and the influence function corresponding to $\tilde{\tau}_{\mathrm{dr}}$ is

$$\tilde{\mathrm{IF}}_\tau(Y, X, D, T) = \frac{1}{q}\left[D\Delta(X) - D\tau + \frac{DT}{p(X)}\{Y - m_1(X)\} - \frac{D(1-T)}{1-p(X)}\{Y - m_0(X)\}\right],$$

see Robins et al. (1994); Hahn (1998); Bang and Robins (2005) for derivations. It can be verified that

$$\left.\frac{\partial \tau(\theta)}{\partial \theta}\right|_{\theta=\theta_0} = E\left\{ \tilde{\mathrm{IF}}_\tau(Y, X, D, T) \cdot S(Y, X, D, T \mid \theta_0) \right\}.$$

Then by the equality (S.2),

$$E\left[\left\{ \mathrm{IF}_\tau(Y, X, D, T) - \tilde{\mathrm{IF}}_\tau(Y, X, D, T) \right\} \cdot S(Y, X, D, T \mid \theta_0)\right] = 0,$$

which implies that $\mathrm{IF}_\tau(Y, X, D, T) - \tilde{\mathrm{IF}}_\tau(Y, X, D, T) \in \mathcal{T}^\perp$.

Note $\mathrm{IF}_\tau(Y,X,D,T) \in \mathcal{T}$, therefore we have

$$E\left[\left\{\tilde{\mathrm{IF}}_\tau(Y,X,D,T) - \mathrm{IF}_\tau(Y,X,D,T)\right\} \cdot \mathrm{IF}_\tau(Y,X,D,T)\right] = 0.$$

The efficiency bounds are the expected squares of the two corresponding influence functions, and their difference is

$$\begin{aligned}
& E\{\tilde{\mathrm{IF}}_\tau(Y,X,D,T)\}^2 - E\{\mathrm{IF}_\tau(Y,X,D,T)\}^2 \\
= \;& E\{\mathrm{IF}_\tau(Y,X,D,T) + \tilde{\mathrm{IF}}_\tau(Y,X,D,T) - \mathrm{IF}_\tau(Y,X,D,T)\}^2 - E\{\mathrm{IF}_\tau(Y,X,D,T)\}^2 \\
= \;& E\{\tilde{\mathrm{IF}}_\tau(Y,X,D,T) - \mathrm{IF}_\tau(Y,X,D,T)\}^2 + 2E[\{\tilde{\mathrm{IF}}_\tau(Y,X,D,T) - \mathrm{IF}_\tau(Y,X,D,T)\} \cdot \mathrm{IF}_\tau(Y,X,D,T)] \\
= \;& E\{\tilde{\mathrm{IF}}_\tau(Y,X,D,T) - \mathrm{IF}_\tau(Y,X,D,T)\}^2,
\end{aligned}$$

and we can calculate that

$$\begin{aligned}
& q^2 \cdot E\left\{\tilde{\mathrm{IF}}_\tau(Y,X,D,T) - \mathrm{IF}_\tau(Y,X,D,T)\right\}^2 \\
= \;& E\left[(1-T)D\left\{\left(\frac{\pi(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)} - \frac{1}{1-p(X)}\right)\{Y - m_0(X)\}\right\}^2\right] \\
& + E\left[(1-D)\left\{\frac{\pi(X)r(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\{Y - m_0(X)\}\right\}^2\right] \\
= \;& E\left[\{1-p(X)\}\pi(X)\left\{\frac{\pi(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)} - \frac{1}{1-p(X)}\right\}^2 \mathrm{var}(Y_0 \mid X, D=1, T=0)\right] \\
& + E\left[\{1-\pi(X)\}\left\{\frac{\pi(X)r(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\right\}^2 \mathrm{var}(Y_0 \mid X, D=0)\right] \\
= \;& E\left[\{1-p(X)\}\pi(X)\left\{\frac{\pi(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)} - \frac{1}{1-p(X)}\right\}^2 \mathrm{var}(Y_0 \mid X, D=1)\right] \\
& + E\left[\{1-\pi(X)\}\left\{\frac{\pi(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\right\}^2 r(X)\mathrm{var}(Y_0 \mid X, D=1)\right] \\
= \;& E\left[\frac{\pi(X)\{1-\pi(X)\}r(X)}{\{1-p(X)\}[\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)]}\mathrm{var}(Y_0 \mid X, D=1)\right] \\
= \;& E\left[\frac{\pi(X)}{1-p(X)} - \frac{\pi^2(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\mathrm{var}(Y_0 \mid X, D=1)\right] \geq 0.
\end{aligned}$$

By $f(X) = f(X \mid D=1)f(D=1)/\pi(X)$, we rewrite the efficiency gain as

$$E\left[\left\{\frac{1}{1-p(X)} - \frac{1}{1-p(X)+\frac{\mathrm{pr}(X|D=0)\mathrm{pr}(D=0)}{\mathrm{pr}(X|D=1)\mathrm{pr}(D=1)}r(X)}\right\}\frac{\mathrm{var}(Y_0 \mid X, D=1)}{\mathrm{pr}(D=1)} \mid D=1\right]. \tag{S.3}$$

Consider an ideal case where $p(X)$ and conditional variances $\mathrm{var}(Y_0 \mid X, D=d)$ for $d \in \{0,1\}$ are constants. For simplicity, we denote the expectation with respect to $f(X \mid D=1)$ by E_1, denote $\mathrm{pr}(X \mid D=0)/\mathrm{pr}(X \mid D=1)$ by $g(X)$, and denote (S.3) by $E_1[\varphi\{g(X)\}]$. Since φ is concave over the range of g in this case, by Jensen's inequality,

$$\text{(S.3)} = E_1[\varphi\{g(X)\}] \leqslant \varphi[E_1\{g(X)\}] = \varphi(1) = E_1\{\varphi(1)\}.$$

Therefore the efficiency gain (S.3) achieves maximum when $\mathrm{pr}(X \mid D=0)/\mathrm{pr}(X \mid D=1) = 1$.

F Proof of proposition 2

We specify the working models $m_1(X;\beta_1)$, $m_0(X;\beta_0)$, $\pi(X;\alpha)$, $p(X;\phi)$ and $r(X;\eta)$ for the outcome means $\{m_1(X), m_0(X)\}$, the propensity scores $\{\pi(X), p(X)\}$ and the variance ratio $r(X)$ respectively. Let $\{\hat{\beta}_1, \hat{\beta}_0, \hat{\alpha}, \hat{\phi}, \hat{\psi}\}$ denote a set of $n^{1/2}$-consistent estimators of these nuisance parameters. If there is no special instructions, let β_1, β_0, α, ϕ and η represent the true value of the parameters when the corresponding working model is correctly specified. Otherwise we use the subscript "*" to denote the probability limit of the parameter in the case of model misspecification.

We let $q = \mathrm{pr}(D=1)$; it can be consistently estimated by $\hat{q} = \hat{E}(D) = n_1/n$.

To facilitate the proof of proposition 2, we first introduce two estimators of τ,

$$\hat{\tau}_{\mathrm{reg}} = \frac{1}{\hat{q}}\hat{E}\left[D\{m_1(X;\hat{\beta}_1) - m_0(X;\hat{\beta}_0)\}\right], \tag{S.4}$$

$$\hat{\tau}_{\mathrm{ipw}} = \frac{1}{\hat{q}}\hat{E}\left[\frac{DT}{p(X;\hat{\phi})}Y - \pi(X;\hat{\alpha})\frac{D(1-T)+(1-D)r(X;\hat{\eta})}{\pi(X;\hat{\alpha})\{1-p(X;\hat{\phi})\}+\{1-\pi(X;\hat{\alpha})\}r(X;\hat{\eta})}Y\right]. \tag{S.5}$$

The following lemma summarizes the properties of the two estimators in consistency.

Lemma 1. *Under assumptions 1–3 and certain regularity conditions described by Newey and McFadden (1994),*

(i). the estimator $\hat{\tau}_{\text{reg}}$ is consistent if the outcome models are correctly specified;

(ii). the estimator $\hat{\tau}_{\text{ipw}}$ is consistent if the propensity score models are correctly specified.

Before giving the proof of lemma 1, we present two useful results.

Lemma 2. *Under assumptions 1–3,*

$$m_0(X) = E(Y \mid X, T = 0) = E(Y_0 \mid X, D = 1) = E(Y_0 \mid X, D = 0).$$

Proof of lemma 2. By the law of total probability,

$$\begin{aligned}
& E(Y \mid X, T = 0) \\
= \; & E(Y \mid X, D = 1, T = 0)\text{pr}(D = 1 \mid X, T = 0) \\
& + E(Y \mid X, D = 0, T = 0)\text{pr}(D = 0 \mid X, T = 0) \\
= \; & E(Y_0 \mid X, D = 1)\text{pr}(D = 1 \mid X, T = 0) \\
& + E(Y_0 \mid X, D = 0)\text{pr}(D = 0 \mid X, T = 0) \\
= \; & E(Y_0 \mid X, D = 1),
\end{aligned}$$

where the second equality follows from assumptions 1–2, and the third follows from assumption 3.

We complete the proof by noting that $E(Y \mid X, D = 1, T = 0) = E(Y_0 \mid X, D = 1)$ under assumptions 1–2. □

Lemma 3. *Under assumptions 1–3, for an arbitrary nonnegative measurable function $h(X)$,*

$$E\left[\pi(X)\frac{D(1-T) + (1-D)h(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}h(X)}Y\right] = qE(Y_0 \mid D = 1).$$

Proof of lemma 3.

$$
\begin{aligned}
& E\left[\pi(X)\frac{D(1-T)+(1-D)h(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}Y\right] \\
= \; & E\left[\pi(X)\frac{\pi(X)\{1-p(X)\}}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}E(Y \mid X, D=1, T=0)\right. \\
& \left.+\pi(X)\frac{\{1-\pi(X)\}h(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}E(Y \mid X, D=0)\right] \\
= \; & E\left\{\pi(X)E(Y_0 \mid X, D=1)\right\} \\
= \; & E\{\pi(X)\} \cdot E\{E(Y_0 \mid X, D=1) \mid D=1\} \\
= \; & qE(Y_0 \mid D=1),
\end{aligned}
$$

where the second equality follows from assumptions 1–3, and the third equality follows from $E\{\pi(X)g(X)\} = E\{q \cdot g(X) \mid D=1\}$ for any integrable function $g(x)$. □

Proof of lemma 1.

(i). Suppose the outcome model $m_1(X;\beta_1)$ and $m_0(X;\beta_0)$ are correctly specified.

Under assumptions 1–3, we obtain

$$
\begin{aligned}
E\{m_0(X;\beta_0) \mid D=1\} &= E\{E(Y_0 \mid X, D=1) \mid D=1\} \\
&= E(Y_0 \mid D=1),
\end{aligned}
$$

where the first equality follows from lemma 2.

Similarly, under assumptions 1,

$$
\begin{aligned}
E\{m_1(X;\beta_1) \mid D=1\} &= E\{E(Y \mid X, D=1, T=1) \mid D=1\} \\
&= E\{E(Y_1 \mid X, D=1) \mid D=1\} \\
&= E(Y_1 \mid D=1).
\end{aligned}
$$

Therefore, we have

$$
\tau = E(Y_1 - Y_0 \mid D=1) = E\{m_1(X;\beta_1) - m_0(X;\beta_0) \mid D=1\}. \tag{S.6}
$$

When the outcome models are correctly specified, the equality (S.6) yields the unbiased estimating equation (S.4) for τ, and then the consistency and asymptotic normality of the estimator $\hat{\tau}_{\text{reg}}$ can be obtained under standard regularity conditions for estimating equations described by Newey and McFadden (1994).

(ii). Suppose that the propensity score models $\pi(X;\alpha)$ and $p(X;\phi)$ are correctly specified, and under certain regular conditions, the estimators $\hat{\alpha}$ and $\hat{\phi}$ are consistent for α and ϕ respectively. Here we do not require the correctness of the specification for $r(X)$, and suppose η^* is the probability limit of $\hat{\eta}$ in the parametric working model $r(X;\eta)$.

Under assumptions 1–3, by lemma 3,

$$E\left[\pi(X;\alpha)\frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}Y\right]=qE(Y_0\mid D=1). \tag{S.7}$$

Under assumptions 1–2, we also have

$$\begin{aligned}E\left\{\frac{DT}{p(X;\phi)}Y\right\}=&E\{\pi(X)E(Y\mid X,D=1,T=1)\}\\ =&E\{\pi(X)E(Y_1\mid X,D=1)\}\\ =&qE(Y_1\mid D=1).\end{aligned} \tag{S.8}$$

Combine the equalities (S.7) and (S.8), we get

$$\tau=\frac{1}{q}E\left[\frac{DT}{p(X;\phi)}Y-\pi(X;\alpha)\frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}Y\right]. \tag{S.9}$$

Therefore, by the equality (S.9), (S.5) is an unbiased estimating equation for τ if the propensity score models are correctly specified, and the result follows. □

After the above preparations, we turn to prove proposition 2.

Case I: Suppose that the propensity score models are correctly specified, while the outcome models can be misspecified.

Lemma 4. *Under assumptions 1–3, for any nonnegative $h(X)$ and integrable $l_0(X)$, $l_1(X)$,*

$$\frac{1}{q}E\left[\frac{DT}{p(X)}l_1(X) - \pi(X)\frac{D(1-T)+(1-D)h(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}l_0(X)\right] = E\{l_1(X) - l_0(X) \mid D=1\}.$$

Proof of lemma 4. The proof is analogous to that of lemma 3 and the equality (S.8). □

We rewrite the doubly robust estimator $\hat{\tau}_{\text{dr}}$ as

$$\begin{aligned} & \hat{\tau}_{\text{ipw}} + \frac{1}{\hat{q}}\hat{E}\left[D\{m_1(X;\hat{\beta}_1) - m_0(X;\hat{\beta}_0)\} - \frac{DT}{p(X;\hat{\phi})}m_1(X;\hat{\beta}_1) + W(X,D,T)m_0(X;\hat{\beta}_0)\right] \\ = \;& \hat{\tau}_{\text{ipw}} + \mathrm{U}_1(X,Y,D,T;\hat{\beta}_1,\hat{\beta}_0,\hat{\alpha},\hat{\phi},\hat{\eta}). \end{aligned}$$

Then by lemma 1, it suffices to show that $\mathrm{U}_1(X,Y,D,T;\hat{\beta}_1,\hat{\beta}_0,\hat{\alpha},\hat{\phi},\hat{\eta})$ converges to zero.

When the propensity score models are correctly specified, under standard regularity conditions for estimating equations, the parameters $(\hat{\alpha},\hat{\phi})$ are consistent for (α,ϕ), and we have $\mathrm{U}_1(X,Y,D,T;\hat{\beta}_1,\hat{\beta}_0,\hat{\alpha},\hat{\phi},\hat{\eta})$ converges in probability to

$$\begin{aligned} & E\{m_1(X;\beta_1^*) - m_0(X;\beta_0^*) \mid D=1\} \\ & -\frac{1}{q}E\left[\frac{DT}{p(X;\phi)}m_1(X;\beta_1^*) - \pi(X;\alpha)\frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}m_0(X;\beta_0^*)\right], \end{aligned}$$

which equals to zero by lemma 4.

Case II: Suppose that the outcome models are correctly specified, while the propensity score models can be misspecified.

Lemma 5. *Under assumptions 1–3, for any integrable functions $g_0(X)$, $g_1(X)$, and $g_3(X)$,*

$$E\Big[DTg_1(X)\{Y - m_1(X)\} + D(1-T)g_2(X)\{Y - m_0(X)\} + (1-D)g_3(X)\{Y - m_0(X)\}\Big] = 0.$$

Proof of lemma 5. We show $E[(1-D)g_3(X)\{Y - m_0(X)\}] = 0$ as an example to illustrate.

$$\begin{aligned} E[(1-D)g_3(X)\{Y - m_0(X)\}] &= E[E[(1-D)g_3(X)\{Y - m_0(X)\} \mid X, D, T]] \\ &= E[(1-D)g_3(X)\{E(Y \mid X, D, T) - m_0(X)\}] \\ &= E[\{1-\pi(X)\}g_3(X)\{E(Y \mid X, D=0, T=0) - m_0(X)\}] \\ &= E[\{1-\pi(X)\}g_3(X)\{m_0(X) - m_0(X)\}] \\ &= 0, \end{aligned}$$

where the fourth equality follows from assumptions 1–3 and lemma 2.

Analogous to the above derivation, one can show that the left two parts are also equal to zero. □

We rewrite the doubly robust estimator $\hat{\tau}_{\text{dr}}$ as

$$\begin{aligned} & \hat{\tau}_{\text{reg}} + \frac{1}{\hat{q}}\hat{E}\left[\frac{DT}{p(X;\hat{\phi})}\{Y - m_1(X;\hat{\beta}_1)\} - W(X,D,T)\{Y - m_0(X;\hat{\beta}_0)\}\right] \\ = \; & \hat{\tau}_{\text{reg}} + \mathrm{U}_2(X, Y, D, T; \hat{\beta}_1, \hat{\beta}_0, \hat{\alpha}, \hat{\phi}, \hat{\eta}). \end{aligned}$$

Then by lemma 1, it suffices to show that $\mathrm{U}_2(X, Y, D, T; \hat{\beta}_1, \hat{\beta}_0, \hat{\alpha}, \hat{\phi}, \hat{\eta})$ converges to zero.

When the outcome models $m_1(X;\beta_1)$ and $m_0(X;\beta_0)$ are correctly specified, under standard regularity conditions for estimating equations, the parameters $(\hat{\beta}_1, \hat{\beta}_0)$ are consistent for (β_1, β_0), and we have $\mathrm{U}_2(X, Y, D, T; \hat{\beta}_1, \hat{\beta}_0, \hat{\alpha}, \hat{\phi}, \hat{\eta})$ converges in probability to

$$\frac{1}{q}E\left[\frac{DT}{p(X;\phi^*)}\{Y - m_1(X;\beta_1)\} - \frac{\pi(X;\alpha^*)\{D(1-T) + (1-D)r(X;\eta^*)\}}{\pi(X;\alpha^*)\{1-p(X;\phi^*)\} + \{1-\pi(X;\alpha^*)\}r(X;\eta^*)}\{Y - m_0(X;\beta_0)\}\right],$$

which equals to zero by lemma 5.

Case III: Both the outcome models and the propensity score models are correctly specified.

Then we could view this as a special case of either **Case I** or **Case II**, and the consistency of the estimator $\hat{\tau}_{\text{dr}}$ for τ still holds under assumptions 1–3.

G Proof of proposition 3

G.1 The efficient influence function for ψ

Because the observed data are the same to that in the proof of proposition 1, and the mean exchangeability for Y_1 imposes no extra restrictions on the observed data, the joint distribution $f(y, x, s, t)$ is unchanged. Following the same approach, we define the regular parametric submodel, and obtain its corresponding score $S(y, x, s, t \mid \theta)$ as well as the tangent space $\mathcal{T}$.

Under assumptions 1–5 and the parametric submodel, the overall average treatment effect among the population corresponding to all observations can be written in the form as

$$\psi(\theta) = \iint y f_1(y \mid x, \theta) f(x, \theta) dy dx - \iint y f_0(y \mid x, \theta) f(x, \theta) dy dx,$$

and $\psi = \psi(\theta_0)$. Then we have

$$\begin{aligned}\left.\frac{\partial \psi(\theta)}{\partial \theta}\right|_{\theta=\theta_0} =& \iint y S_1(y \mid x, \theta_0) f_1(y \mid x) f(x) dy dx - \iint y S_0(y \mid x, \theta_0) f_0(y \mid x) f(x) dy dx \\ &+ \int \{\Delta(x) - \tau\} S_f(x) f(x) dx,\end{aligned}$$

where $\Delta(x) = \int y f_1(y \mid x) dy - \int y f_0(y \mid x) dy$.

Define $\mathrm{IF}_\psi(Y, X, D, T)$ as

$$\begin{aligned}&\Delta(X) - \psi + \frac{DT}{\pi(X)p(X)}\{Y - m_1(X)\} - \frac{D(1-T)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\{Y - m_0(X)\} \\ &- (1-D)\frac{r(X)}{\pi(X)\{1-p(X)\} + \{1-\pi(X)\}r(X)}\{Y - m_0(X)\},\end{aligned}$$

and by calculations similar to the proof of (S.2), under assumptions 1–5 we can verify that,

$$\left.\frac{\partial \psi(\theta)}{\partial \theta}\right|_{\theta=\theta_0} = E\Big\{\mathrm{IF}_\psi(Y, X, D, T) \cdot S(Y, X, D, T \mid \theta_0)\Big\}. \tag{S.10}$$

Also, $\mathrm{IF}_\psi(Y, X, D, T)$ belongs to the tangent space $\mathcal{T}$. Therefore, it is the efficient influence function for the parameter ψ, and the semiparametric efficiency bound is the expected square of $\mathrm{IF}_\psi(Y, X, D, T)$, that is $B_\psi = E\{\mathrm{IF}_\psi(Y, X, D, T)^2\}$.

We give a proof of the difference between asymptotic variances of $\hat{\psi}_{\mathrm{dr}}$ and $\tilde{\psi}_{\mathrm{dr}}$.

The influence function corresponding to $\tilde{\psi}_{\mathrm{dr}}$ is

$$\tilde{\mathrm{IF}}_{\psi}(Y,X,D,T)=\Delta(X)-\psi+\frac{DT}{\pi(X)p(X)}\{Y-m_1(X)\}-\frac{D(1-T)}{\pi(X)\{1-p(X)\}}\{Y-m_0(X)\},$$

and the related derivation can be found in Dahabreh et al. (2019).

Analogous to the illustration in deriving the difference of asymptotic variances between $\hat{\tau}_{\mathrm{dr}}$ and $\tilde{\tau}_{\mathrm{dr}}$, similarly we have

$$E\left[\left\{\tilde{\mathrm{IF}}_{\psi}(Y,X,D,T)-\mathrm{IF}_{\psi}(Y,X,D,T)\right\}\cdot\mathrm{IF}_{\psi}(Y,X,D,T)\right]=0.$$

Therefore, the asymptotic variance difference is

$$\begin{aligned}
&E\{\tilde{\mathrm{IF}}_{\psi}(Y,X,D,T)\}^2-E\{\mathrm{IF}_{\psi}(Y,X,D,T)\}^2\\
&=E\left\{\tilde{\mathrm{IF}}_{\psi}(Y,X,D,T)-\mathrm{IF}_{\psi}(Y,X,D,T)\right\}^2\\
&=E\left[D(1-T)\left\{\left(\frac{1}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}-\frac{1}{\pi(X)\{1-p(X)\}}\right)\{Y-m_0(X)\}\right\}^2\right]\\
&\quad+E\left[(1-D)\left\{\frac{r(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\{Y-m_0(X)\}\right\}^2\right]\\
&=E\left[\pi(X)\{1-p(X)\}\left\{\frac{1}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}-\frac{1}{\pi(X)\{1-p(X)\}}\right\}^2\mathrm{var}(Y_0\mid X,D=1,T=0)\right]\\
&\quad+E\left[\{1-\pi(X)\}\left\{\frac{r(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\right\}^2\mathrm{var}(Y_0\mid X,D=1,T=0)\right]\\
&=E\left[\pi(X)\{1-p(X)\}\left\{\frac{1}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}-\frac{1}{\pi(X)\{1-p(X)\}}\right\}^2\mathrm{var}(Y_0\mid X,D=1)\right]\\
&\quad+E\left[\{1-\pi(X)\}\left\{\frac{1}{\pi(X)\{1-p(X)\}+\{1-\pi(x)\}r(X)}\right\}^2 r(X)\mathrm{var}(Y_0\mid X,D=1)\right]\\
&=E\left[\frac{\{1-\pi(X)\}r(X)[\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)]}{\pi(X)\{1-p(X)\}[\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)]^2}\mathrm{var}(Y_0\mid X,D=1)\right]\\
&=E\left[\frac{\{1-\pi(X)\}r(X)}{\pi(X)\{1-p(X)\}[\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)]}\mathrm{var}(Y_0\mid X,D=1)\right]\\
&=E\left[\frac{1}{\pi(X)\{1-p(X)\}}-\frac{1}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\mathrm{var}(Y_0\mid X,D=1)\right]\geq 0.
\end{aligned}$$

G.2 The efficient influence function for ξ

Under assumptions 1–5 and the parametric submodel, the average treatment effect among the population corresponding to the external data can be written in the form as

$$\xi(\theta) = \frac{\iint y f_1(y \mid x, \theta)\{1-\pi(x,\theta)\}f(x,\theta)dydx - \iint y f_0(y \mid x, \theta)\{1-\pi(x,\theta)\}f(x,\theta)dydx}{\int\{1-\pi(x,\theta)\}f(x,\theta)dx},$$

and $\xi = \xi(\theta_0)$. Thus

$$\begin{aligned}\left.\frac{\partial\xi(\theta)}{\partial\theta}\right|_{\theta=\theta_0} &= \frac{\int\left\{\int yS_1(y \mid x,\theta_0)f_1(y \mid x)dy - \int yS_0(y \mid x,\theta_0)f_0(y \mid x)dy\right\}\{1-\pi(x)\}f(x)dx}{1-q} \\ &\quad -\frac{\int\{\Delta(x)-\tau\}\dot{\pi}(x)f(x)dx}{1-q} + \frac{\int\{\Delta(x)-\tau\}\{1-\pi(x)\}S_f(x)f(x)dx}{1-q},\end{aligned}$$

where $q = \int \pi(x)f(x)dx$, $\Delta(x) = \int yf_1(y \mid x)dy - \int yf_0(y \mid x)dy$.

Define $\mathrm{IF}_\xi(Y, X, D, T)$ as

$$\begin{aligned}\mathrm{IF}_\xi(Y,X,D,T) &= \frac{1}{1-q}\left[(1-D)\Delta(X) - (1-D)\xi + \frac{1-\pi(X)}{\pi(X)}\frac{DT}{p(X)}\{Y-m_1(X)\}\right. \\ &\quad -\frac{D(1-T)\{1-\pi(X)\}}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\{Y-m_0(X)\} \\ &\quad \left.-(1-D)\frac{\{1-\pi(X)\}r(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\{Y-m_0(X)\}\right].\end{aligned}$$

and by calculations similar to the proof of (S.2), under assumptions 1–5 we can verify that

$$\left.\frac{\partial\xi(\theta)}{\partial\theta}\right|_{\theta=\theta_0} = E\Big\{\mathrm{IF}_\xi(Y,X,D,T)\cdot S(Y,X,D,T \mid \theta_0)\Big\}, \tag{S.11}$$

and $\mathrm{IF}_\xi(Y, X, D, T)$ belongs to the tangent space $\mathcal{T}$. Therefore, $\mathrm{IF}_\xi(Y, X, D, T)$ is the efficient influence function for the parameter ξ, and the semiparametric efficiency bound for regular and asymptotic linear estimators of the parameter ξ is the expected square of $\mathrm{IF}_\xi(Y, X, D, T)$, that is $B_\xi = E\{\mathrm{IF}_\xi(Y, X, D, T)^2\}$.

We also give a proof of the difference between asymptotic variances of $\hat{\xi}_{\mathrm{dr}}$ and $\tilde{\xi}_{\mathrm{dr}}$.

The influence function corresponding to $\tilde{\xi}_{\mathrm{dr}}$ is

$$\begin{aligned}\tilde{\mathrm{IF}}_{\xi}(Y,X,D,T) = \frac{1}{1-q}\Bigg[&(1-D)\{\Delta(X)-\xi\} \\ &+\frac{1-\pi(X)}{\pi(X)}\frac{DT}{p(X)}\{Y-m_1(X)\} - \frac{1-\pi(X)}{\pi(X)}\frac{D(1-T)}{1-p(X)}\{Y-m_0(X)\}\Bigg],\end{aligned}$$

and the related derivation can be found in Rudolph and van der Laan (2017); Dahabreh et al. (2020).

By

$$E\left[\left\{\tilde{\mathrm{IF}}_{\xi}(Y,X,D,T)-\mathrm{IF}_{\xi}(Y,X,D,T)\right\}\cdot \mathrm{IF}_{\xi}(Y,X,D,T)\right]=0,$$

we have

$$\begin{aligned}
& (1-q)^2\cdot E\{\tilde{\mathrm{IF}}_{\xi}(Y,X,D,T)\}^2 - E\{\mathrm{IF}_{\xi}(Y,X,D,T)\}^2 \\
= \;& (1-q)^2\cdot E\left\{\tilde{\mathrm{IF}}_{\xi}(Y,X,D,T)-\mathrm{IF}_{\xi}(Y,X,D,T)\right\}^2 \\
= \;& E\left[(1-T)D\left\{\left(\frac{1-\pi(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}-\frac{1-\pi(X)}{\pi(X)\{1-p(X)\}}\right)\{Y-m_0(X)\}\right\}^2\right] \\
& +E\left[(1-D)\left\{\frac{\{1-\pi(X)\}r(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\{Y-m_0(X)\}\right\}^2\right] \\
= \;& E\left[\{1-p(X)\}\pi(X)\left\{\frac{1-\pi(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}-\frac{1-\pi(X)}{\pi(X)\{1-p(X)\}}\right\}^2\mathrm{var}(Y_0\mid X,D=1)\right] \\
& +E\left[\{1-\pi(X)\}\left\{\frac{1-\pi(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\right\}^2 r(X)\mathrm{var}(Y_0\mid X,D=1)\right] \\
= \;& E\left[\frac{\{1-\pi(X)\}^3 r(X)[\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)]}{\pi(X)\{1-p(X)\}[\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)]^2}\mathrm{var}(Y_0\mid X,D=1)\right] \\
= \;& E\left[\frac{\{1-\pi(X)\}^3 r(X)}{\pi(X)\{1-p(X)\}[\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)]}\mathrm{var}(Y_0\mid X,D=1)\right] \\
= \;& E\left[\frac{\{1-\pi(X)\}^2}{\pi(X)\{1-p(X)\}}-\frac{\{1-\pi(X)\}^2}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\mathrm{var}(Y_0\mid X,D=1)\right] \\
\geq \;& 0.
\end{aligned}$$

H Proof of Proposition 4

H.1 The double robustness of $\hat{\psi}_{\text{dr}}$

In addition to the doubly robust estimator $\hat{\psi}_{\text{dr}}$ given in the main text, we introduce the regression based estimator and the inverse probability weighted estimator as following,

$$\hat{\psi}_{\text{reg}} = \hat{E}\left\{m_1(X;\hat{\beta}_1) - m_0(X;\hat{\beta}_0)\right\}, \tag{S.12}$$

$$\hat{\psi}_{\text{ipw}} = \hat{E}\left[\frac{DT}{\pi(X;\hat{\alpha})p(X;\hat{\phi})}Y - \frac{D(1-T)+(1-D)r(X;\hat{\eta})}{\pi(X;\hat{\alpha})\{1-p(X;\hat{\phi})\}+\{1-\pi(X;\hat{\alpha})\}r(X;\hat{\eta})}Y\right]. \tag{S.13}$$

Lemma 6. *Under assumptions 1–5 and certain regularity conditions described by Newey and McFadden (1994),*

(i). the estimator $\hat{\psi}_{\text{reg}}$ is consistent for ψ if the outcome models are correctly specified;

(ii). the estimator $\hat{\psi}_{\text{ipw}}$ is consistent for ψ if the propensity score models are correctly specified.

Proof of lemma 6. (i). Under assumptions 1–3, we obtain

$$\begin{aligned}
E\{m_0(X;\beta_0)\} &= E\{E(Y_0 \mid X, D=1) \mid D=1\}\text{pr}(D=1) + E\{E(Y_0 \mid X, D=0) \mid D=0\}\text{pr}(D=0) \\
&= E(Y_0 \mid D=1)\text{pr}(D=1) + E(Y_0 \mid D=0)\text{pr}(D=0) \\
&= E(Y_0),
\end{aligned}$$

where the first equality follows from lemma 2 and the law of total probability.

Similarly,

$$\begin{aligned}
E\{m_1(X;\beta_1)\} &= E\{E(Y \mid X, D=1, T=1)\} \\
&= E\{E(Y_1 \mid X, D=1)\} \\
&= E\{E(Y_1 \mid X, D=1)\text{pr}(D=1 \mid X) + E(Y_1 \mid X, D=0)\text{pr}(D=0 \mid X)\} \\
&= E\{E(Y_1 \mid X)\} \\
&= E(Y_1),
\end{aligned}$$

where the second equality follows from assumptions 1–2 and the third follows from assumption 4.

Therefore, we have

$$\psi = E(Y_1 - Y_0) = E\{m_1(X;\beta_1) - m_0(X;\beta_0)\}. \tag{S.14}$$

When the outcome models are correctly specified, we could derive a consistent estimate of β_t by solving the corresponding estimation equation, and we denote the estimate as $\hat{\beta}_t$. Together with the consistency of $\hat{\beta}_t$, the equality (S.14) implies that (S.12) is an unbiased estimating equation, and thereby ψ can be consistently estimated by $\hat{\psi}_{\text{reg}}$.

(ii). Under assumptions 1–5, for an arbitrary nonnegative measurable function $h(X)$,

$$\begin{aligned}
&E\left[\frac{D(1-T)+(1-D)h(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}Y\right] \\
&= E\left[\frac{\pi(X)\{1-p(X)\}}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}E(Y \mid X, D=1, T=0)\right. \\
&\quad \left.+\frac{\{1-\pi(X)\}h(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}E(Y \mid X, D=0)\right] \\
&= E\left\{E(Y_0 \mid X, D=1)\right\} \\
&= E\{E(Y_0 \mid X, D=1)\text{pr}(D=1 \mid X) + E(Y_0 \mid X, D=0)\text{pr}(D=0 \mid X)\} \\
&= E\{E(Y_0 \mid X)\} \\
&= E(Y_0),
\end{aligned} \tag{S.15}$$

where the second equality follows from assumptions 1–3, and the third follows from assumption 3.

Similarly, we also have

$$\begin{aligned}
E\left\{\frac{DT}{\pi(X)p(X)}Y\right\} &= E\left\{\frac{\pi(X)p(X)}{\pi(X)p(X)}E(Y \mid X, D=1, T=1)\right\} \\
&= E\{E(Y_1 \mid X, D=1)\} \\
&= E\{E(Y_1 \mid X)\} \\
&= E(Y_1),
\end{aligned} \tag{S.16}$$

where the second equality follows from assumptions 1–2, and the third follows from assumption 4.

Suppose that the propensity score models $\pi(X;\alpha)$ and $p(X;\phi)$ are correctly specified, and under certain regular conditions, the estimators $\hat{\alpha}$ and $\hat{\phi}$ are consistent for α and ϕ respectively. Here we do not require the correctness of the specification for $r(X)$, and suppose η^* is the probability limit of $\hat{\eta}$ in the parametric working model $r(X;\eta)$.

Then by the equalities (S.15) and (S.16), under assumptions 1–5,

$$\psi = E\left[\frac{DT}{\pi(X;\alpha)p(X;\phi)}Y - \frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}Y\right]. \tag{S.17}$$

By the equality (S.17), (S.13) is an unbiased estimating equation for ψ. Therefore, under standard regularity conditions for estimating equations, ψ can be consistently estimated by $\hat{\psi}_{\text{ipw}}$. □

After the above preparations, next we give the proof for the double robustness of $\hat{\psi}_{\text{dr}}$, i.e., under assumptions 1–5 and regularity conditions described in theorems 2.6 and 3.4 of Newey and McFadden (1994), the estimator $\hat{\psi}_{\text{dr}}$ is consistent and asymptotically normal for ψ if either

(i) the propensity score models are correctly specified, or

(ii) the outcome models are correctly specified.

Case I: Suppose that the propensity score models are correctly specified, while the outcome models can be misspecified.

Lemma 7. *Under assumptions 1–5, for any integrable functions $l_0(X)$, $l_1(X)$, and nonnegative measurable function $h(X)$,*

$$E\left[\frac{DT}{\pi(X)p(X)}l_1(X) - \frac{D(1-T)+(1-D)h(X)}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}h(X)}l_0(X)\right] = E\{l_1(X)-l_0(X)\}.$$

Proof of lemma 7. The proof is analogous to that of lemma 3 and the equality (S.8). □

By lemma 7, we have

$$E\left[\frac{DT}{\pi(X;\alpha)p(X;\phi)}m_1(X;\beta_1^*) - \frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}m_0(X;\beta_0^*)\right]$$
$$= E\{m_1(X;\beta_1^*) - m_0(X;\beta_0^*)\}.$$

Therefore, when the propensity scores are correctly specified, $\hat{\psi}_{\mathrm{dr}}$ converges in probability to

$$\begin{aligned}
&E\{m_1(X;\beta_1^*) - m_0(X;\beta_0^*)\} \\
&+E\left[\frac{DT}{\pi(X;\alpha)p(X;\phi)}Y - \frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}Y\right] \\
&-E\left[\frac{DT}{\pi(X;\alpha)p(X;\phi)}m_1(X;\beta_1^*) - \frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}m_0(X;\beta_0^*)\right] \\
&= E\left[\frac{DT}{\pi(X;\alpha)p(X;\phi)}Y - \frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha)\{1-p(X;\phi)\}+\{1-\pi(X;\alpha)\}r(X;\eta^*)}Y\right] \\
&= E(Y_1 - Y_0),
\end{aligned}$$

where the last equality follows from (S.17).

Case II: Suppose that the outcome models are correctly specified, while the propensity score models can be misspecified.

By lemma 5,

$$\begin{aligned}
E\Bigg[&\frac{DT}{\pi(X;\alpha^*)p(X;\phi^*)}\{Y - m_1(X;\beta_1)\} \\
&- \frac{D(1-T)+(1-D)r(X;\eta^*)}{\pi(X;\alpha^*)\{1-p(X;\phi^*)\}+\{1-\pi(X;\alpha^*)\}r(X;\eta^*)}\{Y - m_0(X;\beta_0)\}\Bigg]
\end{aligned}$$

equals zero. Therefore, when the outcome models are correctly specified, the estimator $\hat{\psi}_{\mathrm{dr}}$ converges in probability to

$$E\left\{m_1(X;\beta_1) - m_0(X;\beta_0)\right\},$$

which equals to ψ by (S.14).

Case III: Both the outcome models and the propensity score models are correctly specified.

Then we could view this as a special case of either **Case I** or **Case II**, and the consistency of the estimator $\hat{\psi}_{\mathrm{dr}}$ for ψ still holds under assumptions 1–5.

H.2 The double robustness of $\hat{\xi}_{\text{dr}}$

We can verify that

$$\hat{\xi}_{\text{dr}} = \left(\hat{\psi}_{\text{dr}} - \hat{\tau}_{\text{dr}} \cdot \frac{n_1}{n}\right) \Big/ \frac{n - n_1}{n}.$$

By the double robustness of $\hat{\tau}_{\text{dr}}$ and $\hat{\psi}_{\text{dr}}$, under assumptions 1–5 and certain regularity conditions described by Newey and McFadden (1994), if either

(i) the outcome models $m_t(X; \beta_t)$ for $t \in \{0, 1\}$ are correct, or

(ii) the propensity score models $\pi(X; \alpha)$ and $p(X; \phi)$ are correct,

then $\hat{\psi}_{\text{dr}}$ and $\hat{\tau}_{\text{dr}}$ converge in probability to ψ and τ, respectively.

Because n_1/n is consistent for $\text{pr}(D = 1)$, we have if either (i) or (ii) holds, the estimator $\hat{\xi}_{\text{dr}}$ converges in probability to

$$\frac{\psi - \tau \cdot \text{pr}(D = 1)}{\text{pr}(D = 0)} = \xi,$$

by the Slutsky's theorem. The last equality follows from $\psi = \tau \cdot \text{pr}(D = 1) + \xi \cdot \text{pr}(D = 0)$.

I Discussions on the case where the trial dataset only contains a treated group

The denominator $\pi(X)\{1 - p(X)\} + \{1 - \pi(X)\}r(X)$ in IF_τ is positive as long as $p(X)\pi(X) < 1$, which is weaker than $p(X) < 1$ required by $\tilde{\tau}_{\text{dr}}$. Therefore, although we have required overlap ($0 < p(X) < 1$) in assumption 2, it can be relaxed to allow for $p(X) = 1$ if $\pi(X) < 1$; that is, for the treated units, similar control units only need to exist in either the clinical trial or the external data. Without external control data, lack of overlap in the clinical trial can result in poor finite sample properties (King and Zeng, 2006; Khan and Tamer, 2010). In previous works, trimming methods such as dropping and downweighting units with extreme estimated propensity score values (Crump et al., 2009; Li et al., 2018; Yang and Ding, 2018, etc.) are used to deal with lack of overlap. These methods change the population of interest when data points are dropped or downweighted. However, by incorporating external control data and leveraging the

mean exchangeability assumption, we can still obtain the treatment effect without changing the population of interest.

Consider the extreme case where the trial dataset only contains a treated group. Such a case is often encountered when a control trial is not appropriate to conduct. For instance, Li and Song (2020) consider a rare disease therapy study where the small sample size limits the ability to conduct two-arm trials. In this case, the average treatment effect τ is not identified solely with the trial data, and thus the trial-based doubly robust estimation fails. However, under the mean exchangeability, τ is identified with the aid of external control data. We let $p(X) = 1$ for all X such that $\text{pr}(X \mid D = 1) > 0$, then the full-data doubly robust estimator $\hat{\tau}_{\text{dr}}$ reduces to

$$\frac{n}{n_1}\hat{E}\left\{D\hat{R}_0 - \frac{1-D}{1-\pi(X;\hat{\alpha})}\pi(X;\hat{\alpha})\hat{R}_0\right\}.$$

This estimator is doubly robust against misspecification of $m_0(X;\beta_0)$ and $\pi(X;\alpha)$, and it is locally efficient under assumptions 1, 3 and $\pi(X) < 1$. Therefore, it is preferred compared to the regression-based estimator proposed by Li and Song (2020), which must rest on correct specification of the outcome model.

Now we derive the corresponding efficient influence function. In such a case, the full observational data distribution simplifies to

$$f_1(y \mid x)^d f_0^*(y \mid x)^{1-d}\pi(x)^d[1-\pi(x)]^{1-d}f(x),$$

and thus the corresponding score $S(y, x, d, t \mid \theta)$ reduces to

$$\check{S}(y, x, d, t \mid \theta) = dS_1(y \mid x, \theta) + (1-d)S_0^*(y \mid x, \theta) + \frac{d - \pi(x,\theta)}{\pi(x,\theta)\{1-\pi(x,\theta)\}}\dot{\pi}(x,\theta) + S_f(x,\theta),$$

where $S_f(x,\theta) = \partial \log f(x,\theta)/\partial\theta$, $S_1(y \mid x,\theta) = \partial \log f_1(y \mid x,\theta)/\partial\theta$, $S_0^*(y \mid x,\theta) = \partial \log f_0^*(y \mid x,\theta)/\partial\theta$, and $\dot{\pi}(x,\theta)$ is pathwise derivative for $\pi(x,\theta)$.

By the score, when $p(X) = 1$, the tangent space becomes

$$\check{\mathcal{T}} = \left\{ dS_1(y \mid x) + (1-d)S_0^*(y \mid x) + [d - \pi(x)]a(x) + S(x) \right\},$$

where $S_1(y \mid x)$ satisfies $\int S_1(y \mid x) f_1(y \mid x) dy = 0$, $S_0^*(y \mid x)$ satisfies $\int S_0^*(y \mid x) f_0^*(y \mid x) dy = 0$, $S(x)$ satisfies $\int S(x) f(x) dx = 0$, and $a(x)$ is an arbitrary square-integrable measurable function.

When $p(X) = 1$, the influence function $\mathrm{IF}_\tau(Y, X, D, T)$ reduces to

$$\check{\mathrm{IF}}_\tau(Y, X, D, T) = \frac{1}{q} \left[D\{Y - m_0(X)\} - D\tau + (1-D)\frac{\pi(X)}{1-\pi(X)}\{Y - m_0(X)\} \right].$$

By calculations similar to the proof of (S.2), under assumptions 1, 3 and $\pi(X) < 1$, we can verify that,

$$\left.\frac{\partial \tau(\theta)}{\partial \theta}\right|_{\theta=\theta_0} = E\left\{ \check{\mathrm{IF}}_\tau(Y, X, D, T) \cdot \check{S}(Y, X, D, T \mid \theta_0) \right\}.$$

Also, $\check{\mathrm{IF}}_\tau(Y, X, D, T)$ belongs to the tangent space $\check{\mathcal{T}}$. Therefore, in this case, the reduced form of $\mathrm{IF}_\tau(Y, X, D, T)$, i.e., $\check{\mathrm{IF}}_\tau(Y, X, D, T)$ remains to be the efficient influence function.

J Discussions on the efficiency gain

J.1 An ideal example

As a special case, the following example illustrates the extent of efficiency gain due to external controls when a randomized clinical trial is of interest.

Suppose the trial data are obtained from a randomized clinical trial with a constant treatment propensity score $p(X) = p$. For illustration, we consider an ideal case where $\mathrm{var}(Y_1 \mid X, D = 1) = \mathrm{var}(Y_0 \mid X, D = 1)$, $m_1(X) - m_0(X) = \tau$, $r(X) = r$, and $\pi(X) = \pi$, i.e., the conditional treatment effect, the variance ratio, and the selection propensity score are constants. Then the ratio of B_τ to $\tilde{B}_\tau$ simplifies to

$$B_\tau / \tilde{B}_\tau = 1 - p + \frac{p(1-p)}{1 - p + \frac{1-\pi}{\pi} r}. \tag{S.18}$$

This relative efficiency is monotonically increasing in π and decreasing in r; it can at best achieve $1-p$ as π goes to zero or r goes to infinity, i.e., when the external control dataset is very large or has very small noise.

Moreover, in this ideal case, given π and r, the best p that minimizes B_τ is $\min\{1, 1/2+(1-\pi)r/(2\pi)\}$. This is useful to determine the assignment probability of the active treatment in the trial design, when external data are available from historical databases prior to clinical trials. For $r=1$, when $\pi < 0.5$ the best $p=1$ and otherwise $p=1/(2\pi)$. It suggests that the researchers can assign more units in the clinical trial to receive the active treatment for a higher efficiency as long as more historical control data are available.

J.2 Simulations on the role of sample proportions

We conduct a simulation study to support our claims about the efficiency promotion in the main text. To assess the efficiency gain from the external data, we fix the sample size of the trial dataset, and vary the proportions $\text{pr}(D=1)$ and $\text{pr}(T=1 \mid D=1)$ in the data generating process; a smaller $\text{pr}(D=1)$ means more external control samples are available, while a smaller $\text{pr}(T=1 \mid D=1)$ means less control samples in the trial dataset. We report the ratio of the estimated variances of estimators, $\widehat{\text{var}}(\hat{\tau}_{\text{dr}})/\widehat{\text{var}}(\tilde{\tau}_{\text{dr}})$. We employ the correctly specified outcome models and propensity score models. Both estimators are consistent and asymptotic normal, and their asymptotic variances can also be estimated via the mean square of their influence functions.

Analogous to the setting used in the section 4 of the main text, we generate the baseline covariates $X=(X_1, X_2)$ by a multivariate normal distribution with mean $(0,0)$, variance $(1,1)$ and correlation $\rho = 0.1$. For generating binary variables (D, T), we consider the following strategy,

$$\begin{cases} D \mid X \sim \text{Ber}([1+\exp\{-(a+0.2X_1-0.2X_2)\}]^{-1}), \\ T \mid (X, D=1) \sim \text{Ber}([1+\exp\{-(b+0.2X_1+0.3X_2)\}]^{-1}), \end{cases}$$

where the parameters a and b can be tuned to change the proportions $\text{pr}(D=1)$ and $\text{pr}(T=1 \mid D=1)$ to a user-specified level. By tuning the parameters a and b, we could obtain the

corresponding $\pi(X) = \text{pr}(D = 1 \mid X)$ and $p(X) = \text{pr}(T = 1 \mid X, D = 1)$ such that $E\{\pi(X)\} = q$ and $E\{p(X) \mid D = 1\} = p$ achieve the user-specified levels. In this numerical study, we consider the scenarios where the proportion of external data $1 - q$ is at the levels of 0.3, 0.4, 0.5, 0.6 and 0.7, and the treated group proportion within the trial data p is at the levels of 0.5 and 0.7, respectively.

For the observed outcomes Y, we consider the following strategy to generate,

$$\begin{cases} Y \mid (X, D = 1, T = 1) \sim N(3 + X_1 + 1.5X_2,\ 1), \\ Y \mid (X, D = d, T = 0) \sim N(1 + X_1 + X_2,\ 1), \quad d \in \{0, 1\}. \end{cases}$$

Throughout the simulation, the sample size of the trial dataset is kept around 500 and we replicate 1000 times for each scenario.

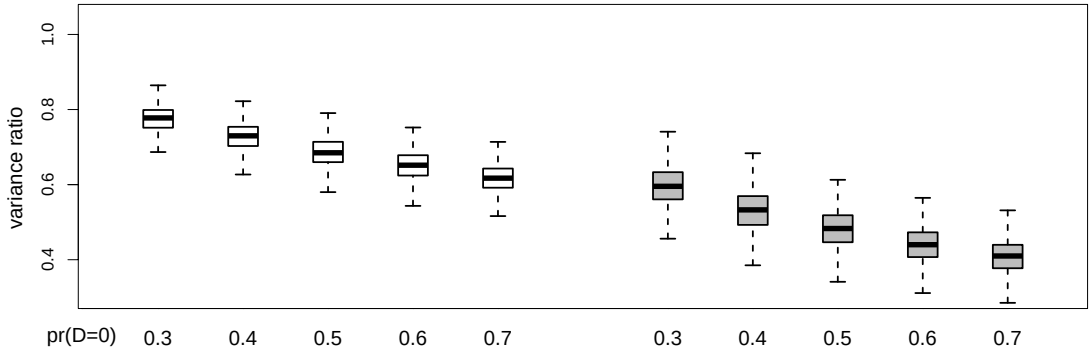

Figure J1: Boxplots of $\widehat{\text{var}}(\hat{\tau}_{\text{dr}})/\widehat{\text{var}}(\tilde{\tau}_{\text{dr}})$, the ratio of the estimated variances. White boxes are for $\text{pr}(T = 1 \mid D = 1) = 0.5$ and grey ones for 0.7.

Fig. J1 shows the results. All variance ratios under different proportions $\text{pr}(D = 1)$ and $\text{pr}(T = 1 \mid D = 1)$ are remarkably smaller than one, which is evidence that the full-data estimator is more efficient. On average, the variance ratios decrease as the proportion of external data and the proportion of the treated group in the trial data increase. These results inspire researchers that it is more crucial to incorporate external control data into a trial dataset with fewer control units.

J.3 Simulations on the role of $r(X)$

In addition, we do more simulations to investigate the role of the variance ratio in efficiency improvement. As analyzed in the main text, the variance ratio can be regarded as a weight tuning function, which strikes a balance between the control units in the external and external data. A larger $r(X)$ allows us to attach more importance to the external control data compared to the control group in the trial data. For illustration purpose, we consider a constant variance ratio. We use the data generating mechanism of X described previously, and generate (D, T, Y) by the following mechanism:

$$\begin{cases} D \mid X \sim \text{Ber}([1+\exp\{-(3+0.2X_1-0.2X_2)\}]^{-1}), \\ T \mid (X, D=1) \sim \text{Ber}([1+\exp\{-(0.2X_1+0.3X_2)\}]^{-1}); \end{cases}$$

$$\begin{cases} Y \mid (X, D=1, T=1) \sim N\left(3+0.5X_1+1.5X_2,\ 1\right), \\ Y \mid (X, D=d, T=0) \sim N\left(1+0.5X_1+X_2,\ 1+9d\right); \end{cases}$$

where $d \in \{0, 1\}$. In this case, the true value of the variance ratio $r(X)$ is 10.

We vary the value of $r(X)$ used in the full-data doubly robust estimator $\hat{\tau}_{\text{dr}}$. Besides, we employ the correct parametric models for the outcome means and propensity scores. We simulate 1000 replicates under 1000 sample size and summarize the results with bias boxplots in Fig. J2.

As expected, when the $r(X)$ takes its true value 10, the estimator $\hat{\tau}_{\text{dr}}$ achieves the best precision with the smallest variability. Too large or too small values of $r(X)$ can affect the efficiency of the full-data doubly robust estimator, but not the consistency. Moreover, when $r(X) = 0$, the external control data are ignored, and therefore the estimator $\hat{\tau}_{\text{dr}}$ has a same performance to the trial-based doubly robust estimator $\tilde{\tau}_{\text{dr}}$ in this case. However, when we set $r(X) > 0$, the full-data doubly robust estimator would always have a smaller variability.

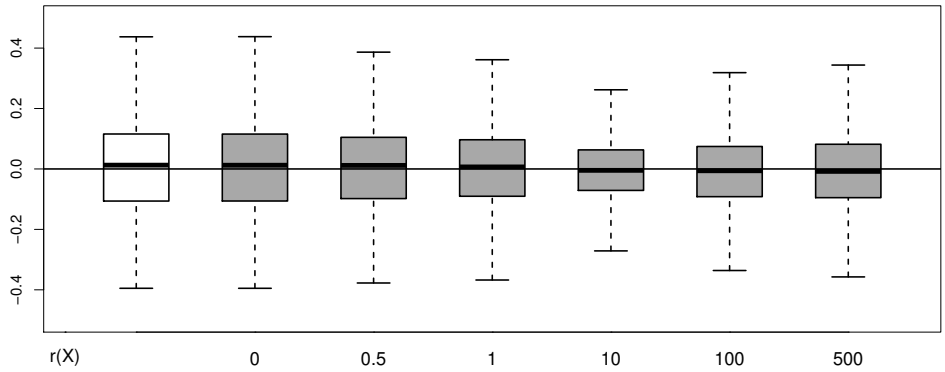

Figure J2: Bias boxplots for estimators of the treatment effect τ. The white box is for the trial-based doubly robust estimator $\tilde{\tau}_{\mathrm{dr}}$, and the grey ones are for the full-data doubly robust estimator $\hat{\tau}_{\mathrm{dr}}$ using different values of $r(X)$.

K Simulation settings for section 4 of the main text

We generate the baseline covariates $X = (X_1, X_2)$ by a multivariate normal distribution with mean $(0, 0)$, variance $(1, 1)$ and correlation $\rho = 0.1$. For generating binary variables (D, T), we consider the following two strategies using the bernoulli distribution,

$$\begin{cases} D \mid X \sim \mathrm{Ber}([1+\exp\{-(0.2X_1 - 0.2X_2)\}]^{-1}), \\ T \mid (X, D=1) \sim \mathrm{Ber}([1+\exp\{-(0.2X_1 + 0.3X_2)\}]^{-1}); \end{cases}$$

$$\begin{cases} D \mid X \sim \mathrm{Ber}([1+\exp\{-(0.2X_1 - 0.1X_1^2 + 0.3X_2^2)\}]^{-1}), \\ T \mid (X, D=1) \sim \mathrm{Ber}([1+\exp\{-(0.2X_1 + 0.3X_2 + \mathrm{sign}(-\mathrm{X}_2)\cdot 1.5\mathrm{X}_2^2)\}]^{-1}). \end{cases}$$

For these settings, the missing data proportions $\mathrm{pr}(D=1)$ and $\mathrm{pr}(T=1 \mid D=1)$ are about 50%. Similarly, we also consider two strategies to generate the observed outcomes Y with the normal distribution,

$$\begin{cases} Y \mid (X, D=1, T=1) \sim N\left(3 + 0.5X_1 + 1.5X_2,\ 0.2|X_2|^{0.2}\right), \\ Y \mid (X, D=d, T=0) \sim N\left(1 + 0.5X_1 + X_2,\ \sigma_d^2(X)\right); \end{cases}$$

$$\begin{cases} Y \mid (X, D=1, T=1) \sim N\left(3 - 0.2X_1^2 - 0.4X_2^2,\ 0.2|X_2|^{0.2}\right), \\ Y \mid (X, D=d, T=0) \sim N\left(1 + 0.6X_1^2 + 0.6X_2^2,\ \sigma_d^2(X)\right); \end{cases}$$

where $d \in \{0, 1\}$, $\sigma_1^2(X) = 2|X_2|^{0.2}$ and $\sigma_0^2(X) = |X_1|^{0.4}$.

In particular, we design four scenarios: (i) generating (D, T) and Y both by the first strategy respectively; (ii) generating Y by the first strategy but (D, T) by the second; (iii) generating Y by the second strategy but (D, T) by the first; (iv) generating (D, T) and Y both by the second strategy respectively.

We employ the parametric models for estimation such that both the outcome models and propensity score models are correctly specified under the first scenario. Therefore, the propensity score models and the outcome models would be misspecified under the second scenario and the third scenario respectively. While under the fourth scenario, both the parametric models are misspecified. We simulate 1000 replicates under 1000 sample size for each scenario and summarize the results for estimating τ in the main text.

L Simulation results for treatment effects ψ and ξ

We present simulation results about the performance of estimators for ψ and ξ. The data generating mechanisms and the specification of working models are same to that described in section K. We simulate 1000 replicates under 1000 sample size for each scenario and summarize the results with bias boxplots in Fig. L1 for ψ and Fig. L2 for ξ. Table L1 shows coverage probabilities of the 95% confidence interval based on normal approximation.

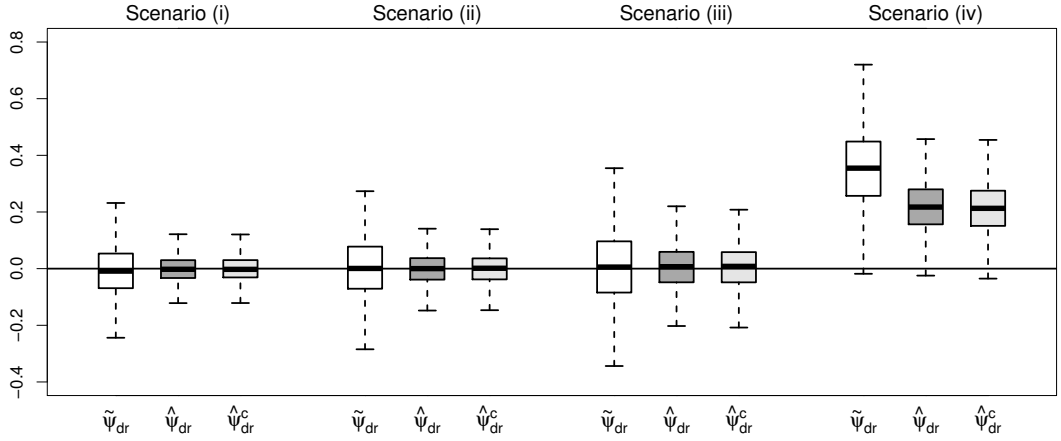

Figure L1: Bias boxplots for estimators of the treatment effect ψ.

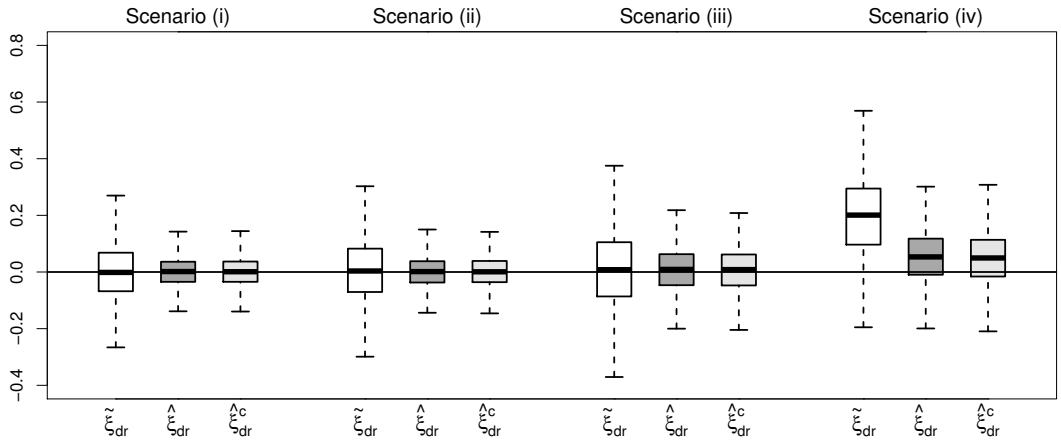

Figure L2: Bias boxplots for estimators of the treatment effect ξ.

Table L1: Coverage probability of 95% confidence interval

Scenario	ψ			ξ		
	$\tilde{\psi}_{\mathrm{dr}}$	$\hat{\psi}_{\mathrm{dr}}$	$\hat{\xi}^{c}_{\mathrm{dr}}$	$\tilde{\xi}_{\mathrm{dr}}$	$\hat{\xi}_{\mathrm{dr}}$	$\hat{\psi}^{c}_{\mathrm{dr}}$
(i)	0.947	0.948	0.951	0.944	0.957	0.957
(ii)	0.944	0.956	0.956	0.948	0.958	0.960
(iii)	0.957	0.950	0.951	0.955	0.942	0.946
(iv)	0.314	0.332	0.314	0.711	0.866	0.892

Note for Fig. L1, L2 and Table L1: $\hat{\psi}^{c}_{\mathrm{dr}}$ and $\hat{\xi}^{c}_{\mathrm{dr}}$ are obtained by using using a constant variance ratio in the full-data doubly robust estimators of ψ and ξ, respectively.

References

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

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2020). Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047* .

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96,** 187–199.

Dahabreh, I. J., Haneuse, S. J. A., Robins, J. M., Robertson, S. E., Buchanan, A. L., Stuart, E. A., and Hernán, M. A. (2021). Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology* **190,** 1632–1642.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine* **39,** 1999–2014.

Dahabreh, I. J., Robertson, S. E., Tchetgen Tchetgen, E., Stuart, E. A., and Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75,** 685–694.

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

Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica* **78,** 2021–2042.

King, G. and Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis* **14,** 131–159.

Li, F., Buchanan, A. L., and Cole, S. R. (2021). Generalizing trial evidence to target populations in non-nested designs: Applications to aids clinical trials. *arXiv preprint arXiv:2103.04907* .

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

Li, X. H. and Song, Y. (2020). Target population statistical inference with data integration across multiple sources—an approach to mitigate information shortage in rare disease clinical trials. *Statistics in Biopharmaceutical Research* **12,** 322–333.

Luedtke, A., Carone, M., and van der Laan, M. J. (2019). An omnibus non-parametric test of equality in distribution for unknown functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81,** 75–99.

Newey, K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics (Edited by R. Engle and D. McFadden)*, volume 4, pages 2112–2245. New York: Elsevier Science.

Olschewski, M. and Scheurlen, H. (1985). Comprehensive cohort study: An alternative to randomized consent design in a breast preservation trial. *Methods of Information In medicine* **24,** 131–134.

Pearl, J., Bareinboim, E., et al. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science* **29,** 579–595.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Rudolph, K. E. and van der Laan, M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 1509.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., and Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology* **186,** 1010–1014.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* **105,** 487–493.