A Comparison of Methods for Estimating the Average Treatment Effect on the Treated for Externally Controlled Trials

Huan WANG[1], Fei WU[1], and Yeh-Fong CHEN*

Abstract

While randomized trials may be the gold standard for evaluating the effectiveness of the treatment intervention, in some special circumstances, single-arm clinical trials utilizing external control may be considered. The causal treatment effect of interest for single-arm trials is usually the average treatment effect on the treated (ATT) rather than the average treatment effect (ATE). Although methods have been developed to estimate the ATT, the selection and use of these methods require a thorough comparison and in-depth understanding of the advantages and disadvantages of these methods. In this study, we conducted simulations under different identifiability assumptions to compare the performance metrics (e.g., bias, standard deviation (SD), mean squared error (MSE), type I error rate) for a variety of methods, including the regression model, propensity score matching (PSM), Mahalanobis distance matching (MDM), coarsened exact matching, inverse probability weighting, augmented inverse probability weighting (AIPW), AIPW with SuperLearner, and targeted maximum likelihood estimator (TMLE) with SuperLearner.

Our simulation results demonstrate that the doubly robust methods in general have smaller biases than other methods. In terms of SD, nonmatching methods in general have smaller SDs than matching-based methods. The performance of MSE is a trade-off between the bias and SD, and no method consistently performs better in term of MSE. The identifiability assumptions are critical to the models' performance: Violation of the positivity assumption can lead to a significant inflation of type I errors in some methods; violation of the unconfoundedness assumption can lead to a large bias for all methods.

According to the simulation results, under most scenarios we examined, PSM and MDM methods perform best overall in terms of type I error control. However, they in general have worse performance in the estimation accuracy compared to doubly robust methods given that the identifiability assumptions are not severely violated.

KEYWORDS AND PHRASES: Single-arm trials, External control, Real-world data, Causal inference, Average treatment effect on the treated.

1. INTRODUCTION

Randomized clinical trials (RCTs) are usually considered the gold standard for demonstrating the efficacy and safety of an intervention. However, for rare diseases it is difficult to recruit a large number of patients, and for severe diseases it is unethical to assign patients to placebo. Single-arm trials utilizing an external real-world data (RWD) control could be an appealing option. In addition to the feasibility and ethical advantages, single-arm trials with external RWD controls may save time and money and as a result have been receiving increasing attention.

To draw inferences about the effect of a treatment, a target causal effect needs to be defined. For the single-arm trials with external RWD controls, the causal treatment effect of interest could be the average treatment effect on the treated (ATT) (i.e., the average treatment effect for the population receiving the treatment compared to what would have happened if they had not received treatment)

if the treatment effect among the enrolled population is of more interest [16, 17]. This is particularly the case if there is heterogeneity between patients participating and those not participating in the trial due to eligibility criteria. For example, if the experimental drug is a second-line treatment, the study drug's effect cannot be considered for naive patients who are not eligible to participate in the trial.

The ATT can also be of interest if there is heterogeneity in the baseline characteristics between patients participating and those who do not. E.g., patients who participate in the trial are older than patients who do not. In this study, we focus on estimating the causal effect of ATT.

Many well-established methods have been proposed to estimate ATT, including the coarsened exact matching (CEM), propensity score matching (PSM), Mahalanobis distance matching (MDM), inverse propensity weighting (IPW), augmented inverse propensity weighting (AIPW), targeted maximum likelihood estimator (TMLE), etc. Attempts have been made to compare the performance of these models in terms of estimating ATT. For example, Abdia et al. compared the performance of the PSM, regression model,

---

*Corresponding author.
[1]Contributed equally.

stratification model, and IPW [2]; Chatton et al. compared the performance of the g-computation method, IPW, PSM, and TMLE with different covariates sets [9]. However, the choice of the model remains unclear and controversial in the actual design stage of single-arm clinical trials with RWD controls.

In this study, we examined the performance of 10 methods for causal inference, including the regression model, PSM, MDM, CEM, IPW, AIPW, AIPW with SuperLearner, and TMLE with SuperLearner via simulations. The metrics of performance used for comparison included bias, standard deviation (SD), mean squared error (MSE), and type I error. With our simulation results, practical advice on the use of these methods in single-arm clinical trials with RWD controls is provided.

This paper is organized as follows. The introduction section provides background information on single-arm trials with external RWD controls and presents the motivation behind this study. The Methods section revisits the definitions of ATT and identifiability assumptions and provides a description of the models and simulation settings utilized in this study. The Results section presents the findings from various scenarios. In the Discussion section, we analyze the experimental results and provide recommendations for planning single-arm trials with external controls. Finally, the Conclusion section summarizes the study's key findings, interpretation, and implications.

2. METHODS

2.1 Average Treatment Effect on the Treated

In a real-world study, a causal effect of a treatment can be defined in terms of the comparison between potential outcomes (also known as the counterfactual outcomes) of a population who would receive the treatment and potential outcomes of the same population who would not receive the treatment [18]. The comparison is drawn from a population level instead of an individual level because only one of the potential outcomes can be observed for each patient [1]. In a nonrandomized, single-arm clinical trial with an external RWD control, the population of interest is usually only the patients receiving the treatment [13]. Correspondingly, the causal effect of interest is the ATT.

For patient with index i, let Z_i denote the patient's treatment assignment with a value of 1 for treated and 0 for not treated, $\boldsymbol{X}_i$ vector of the patient's covariates, Y_i the patient's observed outcome, $Y_i(1)$ and $Y_i(0)$ the patient's potential outcomes corresponding to $Z_i = 1$ and $Z_i = 0$, respectively. Then the ATT of our interest in terms of the difference scale τ_{ATT} rather than others (e.g., ratio), can be formally defined as

$$\tau_{\text{ATT}} \equiv E[Y_i(1) - Y_i(0)|Z_i = 1].$$

2.2 Identifiability Assumptions

An average causal effect such as ATT is not always estimable. Hernán and Robins summarized three assumptions needed to estimate an average causal effect: consistency, positivity, and conditional exchangeability [18]. The definitions of these assumptions are as follows.

1. Consistency: $Y_i = Y_i(1)$ if $Z_i = 1$ and $Y_i = Y_i(0)$ if $Z_i = 0$.
2. Positivity: $0 < P(Z_i = 1|\boldsymbol{X_i} = \boldsymbol{x_i}) < 1$ for any $\boldsymbol{X_i}$ with $P(\boldsymbol{X_i} = \boldsymbol{x_i}) > 0$.
3. Conditional exchangeability (unconfoundedness): $Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i||\boldsymbol{X}_i$.

The consistency assumption means that a subject's potential outcome under his or her observed exposure history is equal to the subject's observed outcome. The consistency assumption requires that the treatment should be sufficiently well-defined, contingent upon agreement among experts based on the available substantive knowledge [18].

The positivity assumption implies that all subjects have a non-zero probability of receiving (or not receiving) the treatment. Positivity violations can be divided into two categories: 1) theoretical violation that the probability of being treated is 0 or 1, e.g., certain patients are restricted from receiving the treatment or control, and 2) near or practical violation in the sampling due to high/low exposure prevalence or small sample size [26]. One method for assessing the positivity assumption is via the propensity score (PS), which is the probability of receiving the treatment conditional on covariate values, i.e., $PS(\boldsymbol{X}_i) = P(Z_i = 1|\boldsymbol{X}_i)$ [38]. In particular, we can study the distribution of the PS for the study population and explore whether there are subjects with PS values near 0 or 1 (Figure 1). If many PS values are distributed around 0 or 1, then the positivity assumption should not hold; on the other hand, if no PS values are distributed around 0 or 1, the positivity assumption can be considered reasonable.

The conditional exchangeability assumption, also known as the unconfoundedness assumption, suggests that all variables influencing treatment selection and outcomes are measured and accessible. Violation of this assumption occurs when unmeasured confounders are not adequately accounted for. The unconfoundedness assumption is inherently untestable. However, many sensitivity analyses have been developed to indirectly assess the unconfoundedness assumption, i.e., assessing the bias of the causal effect estimate by assuming unmeasured confounding [30, 24, 11, 36].

2.3 Methods for Causal Inference

2.3.1 Linear Regression Model

In general, a regression model is not a sound model for estimating ATT. However, if the true outcome model can be properly fitted by a regression model, the regression model may estimate ATT well. In particular, if 1) the identifiability

A

Possibility of positivity violation: low

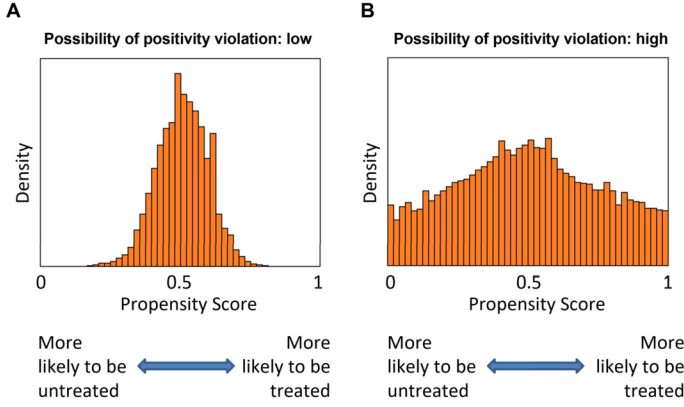

B

Possibility of positivity violation: high

Figure 1: Distributions of PS values in study populations characterized by low (A) and high (B) potential for violating the positivity assumption. In these populations, subjects with PS values close to 0 or 1 are more prone to violating the positivity assumption.

assumptions are met and 2) the outcome model is correctly specified, the ATT can be estimated through the regression model. However, in practice, it is challenging to specify the correct outcome model. In this study, we use the linear regression (LR) that includes the specified covariates $\boldsymbol{X}_i$ and the treatment variable Z_i without interaction terms as a simple reference model. The ATT is estimated through the coefficient of the treatment variable Z_i.

2.3.2 Matching

A fundamental problem of causal inference is that we can only observe one potential outcome from each observation. However, the information from the potential outcome that we do not observe is needed in the estimation of the causal effect. An intuitive approach is to estimate the unobserved potential outcome with a similar one via matching. An underlying assumption about the matching is that observations similar on their covariate values are also similar on the potential outcomes [25]. For estimating ATT using matching, we can assume that the expectation of $Y_i(0)$ and $Y_i(0)^M$ are the same where $Y_i(0)^M$ is the sample in the control group that matches $Y_i(1)$ on $\boldsymbol{X}_i$, i.e.,

$$E[Y_i(0)|\boldsymbol{X}_i] = E[Y_i(0)^M|\boldsymbol{X}_i]. \tag{2.1}$$

Then by the unconfoundedness assumption, ATT can be expressed as

$$
\begin{aligned}
\tau_{\text{ATT}} =& E[Y_i(1) - Y_i(0)|Z_i = 1] \\
=& E_{\boldsymbol{X}_i|Z_i=1}\{E[Y_i(1)|\boldsymbol{X}_i] - E[Y_i(0)|\boldsymbol{X}_i]\} \\
=& E_{\boldsymbol{X}_i|Z_i=1}\{E[Y_i(1)|\boldsymbol{X}_i] - E[Y_i(0)^M|\boldsymbol{X}_i]\} \\
=& E[Y_i(1) - Y_i(0)^M|Z_i = 1]
\end{aligned}
$$

and estimated by

$$\frac{1}{N_1} \sum_{i \in [Z=1]} [Y_i(1) - Y_i(0)^M],$$

where N_1 is the sample size of the treatment group.

2.3.3 Coarsened Exact Matching

In practice, exact matching of covariates may not be feasible when there are continuous covariates or when there are categorical covariates with many levels. One approach to address this issue is the CEM method, which is a form of exact matching based on the subclasses of covariates [22]. It involves two steps: 1) coarsening each covariate into several "bins," i.e., subclasses, and 2) performing exact matching on those bins. The unmatched units are discarded.

For CEM, the choice of the number of bins is crucial to model performance. Domain knowledge is required to determine the appropriate number of bin for each variable [22]. In this study, we arbitrarily choose 2 (denoted by CEM2), a smaller number of bin, and 5 (denoted by CEM5), a reasonably larger one, for all covariates to study the performance of the CEM model with different numbers of bins. For each number of bins, the boundaries of the bins were set to be evenly spaced between the maximum and minimum values of the covariate [20].

2.3.4 Mahalanobis Distance Matching

An alternative way to loosen the exact match requirement is by using MDM so that similar subjects are matched. MDM pairs subjects that are close based on the Mahalanobis distance. The Mahalanobis distance between Patients u and v, denoted by $d_{u,v}$, is defined as

$$d_{u,v} = \sqrt{(\boldsymbol{X}_u - \boldsymbol{X}_v)^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_u - \boldsymbol{X}_v)},$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the covariates.

2.3.5 Propensity Score Matching

Matching can also be conducted on the PS of the relevant covariates instead of on the covariates themselves. Rosenbaum and Rubin proved that PS is a balancing score and that potential outcomes are independent of treatment conditional on a balancing score if the unconfoundedness assumption holds [31]. Then the assumption of unconfoundedness for the given covariates can be relaxed to be unconfoundedness for the given PS values, i.e.,

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid X_i \ \Rightarrow Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid PS(\boldsymbol{X}_i).$$

In addition, according to equation (2.1) and the law of total expectation, we have

$$
\begin{aligned}
E[Y_i(0)|PS(\boldsymbol{X}_i)] =& E\{E[Y_i(0)|\boldsymbol{X}_i]|PS(\boldsymbol{X}_i)\} \\
=& E\{E[Y_i(0)^{PSM}|\boldsymbol{X}_i]|PS(\boldsymbol{X}_i)\}
\end{aligned}
$$

$$=E[Y_i^{PSM}(0)|PS(\boldsymbol{X}_i)], \qquad (2.2)$$

where $Y_i(0)^{PSM}$ is the sample in the control group that matches $Y_i(1)$ on $PS(\boldsymbol{X}_i)$. Then, by the assumption of the unconfoundedness given the PS and equation (2.2), ATT can be obtained by

$$
\begin{aligned}
\tau_{\text{ATT}} =& E[Y_i(1) - Y_i(0)|Z_i = 1] \\
=& E_{PS(\boldsymbol{X}_i)|Z_i=1}\{E[Y_i(1)|PS(\boldsymbol{X}_i)] - E[Y_i(0)|PS(\boldsymbol{X}_i)]\} \\
=& E_{PS(\boldsymbol{X}_i)|Z_i=1}\{E[Y_i(1)|PS(\boldsymbol{X}_i)] \\
& - E[Y_i(0)^{PSM}|PS(\boldsymbol{X}_i)]\} \\
=& E[Y_i(1) - Y_i(0)^{PSM}|Z_i = 1]
\end{aligned}
$$

and estimated by

$$\frac{1}{N_1}\sum_{i \in [Z=1]} [Y_i(1) - Y_i(0)^{PSM}].$$

For a comprehensive introduction of matching-based methods, including PSM, refer to the overview by Stuart [33].

2.3.6 Inverse Probability Weighting

The IPW is another PS-based method that can be used to estimate the ATT. The IPW uses the PS to balance the characteristics of covariates in the treated and control groups by weighting each subject in the analysis by the inverse probability of receiving his/her actual exposure [10]. In this study, we used a generalized version of the IPW estimator, which has the following expression:

$$\frac{\sum_{i=1}^n W_i Z_i Y_i}{\sum_{i=1}^n W_i Z_i} - \frac{\sum_{i=1}^n W_i(1-Z_i)Y_i}{\sum_{i=1}^n W_i(1-Z_i)}$$

where $W_i = \frac{\widehat{PS}(\boldsymbol{X}_i)}{Z_i\widehat{PS}(\boldsymbol{X}_i)+(1-Z_i)(1-\widehat{PS}(\boldsymbol{X}_i))}$ [27, 28]. This IPW estimator is a consistent estimator of ATT when the identifiability assumptions hold and $\widehat{PS}(\boldsymbol{X}_i)$ is a consistent estimator of the PS.

2.3.7 AIPW

The AIPW method combines the IPW method and an outcome estimation model. It is a doubly robust estimator in the sense that it is consistent if either the treatment assignment model or the potential outcome model is correctly specified, but not necessarily both [29]. In this study, we used the following AIPW estimator for ATT estimation:

$$
\begin{aligned}
&\frac{\sum_{i=1}^n \widehat{PS}(\boldsymbol{X}_i))(\widehat{Q}_1(\boldsymbol{X}_i)) - \widehat{Q}_0(\boldsymbol{X}_i))}{\sum_{i=1}^n \widehat{PS}(\boldsymbol{X}_i)} \\
&+ \frac{\sum_{i=1}^n W_i Z_i (Y_i - \widehat{Q}_1(\boldsymbol{X}_i))}{\sum_{i=1}^n W_i Z_i} \\
&- \frac{\sum_{i=1}^n W_i(1-Z_i)(Y_i - \widehat{Q}_0(\boldsymbol{X}_i))}{\sum_{i=1}^n W_i(1-Z_i)}
\end{aligned}
$$

where $W_i = \frac{\widehat{PS}(\boldsymbol{X}_i)}{Z_i\widehat{PS}(\boldsymbol{X}_i)+(1-Z_i)(1-\widehat{PS}(\boldsymbol{X}_i))}$ and $\widehat{Q}_1(\boldsymbol{X}_i)$ and $\widehat{Q}_0(\boldsymbol{X}_i)$ denote the estimators for $E[Y_i|\boldsymbol{X}_i, Z_i = 1)]$ and $E[Y_i|\boldsymbol{X}_i, Z_i = 0)]$, respectively [28].

2.3.8 Targeted Maximum Likelihood Estimator

The TMLE, introduced by van der Laan and Rubin, is an alternative doubly robust estimation method that is maximum-likelihood–based and includes a targeting step that optimizes the bias-variance tradeoff for the parameter of interest [34, 32].

The application of TMLE for estimating ATT involves the following steps [35, 12]:

1. Obtain an initial estimate of the expected outcome $\widehat{E}^0[Y_i|Z_i, \boldsymbol{X}_i]$ by fitting an outcome model.
2. Obtain an initial estimate of the PS, i.e., $\widehat{PS}^0(Z_i|\boldsymbol{X}_i)$, by fitting a treatment assignment model.
3. Apply the targeted learning theory to update the initial estimates of $\widehat{E}^0[Y_i|Z_i, \boldsymbol{X}_i]$ and $\widehat{PS}^0(Z_i|\boldsymbol{X}_i)$ to $\widehat{E}^*[Y_i|Z_i, \boldsymbol{X}_i]$ and $\widehat{PS}^*(Z_i|\boldsymbol{X}_i)$, respectively. These updates are performed such that they solve the efficient score equation:

$$\sum_{i=1}^n D\left(\widehat{E}^*[Y_i|Z_i, \boldsymbol{X}_i], \widehat{PS}^*(Z_i|\boldsymbol{X}_i), \widehat{P}^*_{\boldsymbol{X}_i|Z_i=1}\right) = 0$$

where D is defined as

$$
\begin{aligned}
&D\left(E[Y_i|Z_i, \boldsymbol{X}_i], PS(Z_i|\boldsymbol{X}_i), P_{\boldsymbol{X}_i|Z_i=1}\right) \\
&= \left[\frac{I(Z_i = 1)}{P(Z_i = 1)} - \frac{I(Z_i = 0)PS(Z_i = 1|\boldsymbol{X}_i)}{P(Z_i = 1)PS(Z_i = 0|\boldsymbol{X}_i)}\right] \\
&\quad \times (Y_i - E[Y_i|Z_i, \boldsymbol{X}_i]) \\
&\quad + \frac{I(Z_i = 1)}{P(Z_i = 1)}(E[Y_i|Z_i = 1, \boldsymbol{X}_i] - E[Y_i|Z_i = 0, \boldsymbol{X}_i] \\
&\qquad\qquad - \tau_{\text{ATT}})
\end{aligned}
$$

and $\widehat{P}^*_{\boldsymbol{X}_i|Z_i=1}$ is the empirical estimator of $P_{\boldsymbol{X}_i|Z_i=1}$.

4. The TMLE estimator for the ATT is then obtained as:

$$
\begin{aligned}
&\widehat{\tau}^*_{\text{ATT}} \\
&= \int \left(\widehat{E}^*[Y_i|Z_i=1, \boldsymbol{X}_i] - \widehat{E}^*[Y_i|Z_i=0, \boldsymbol{X}_i]\right) d\widehat{P}^*_{\boldsymbol{X}_i|Z_i=1}.
\end{aligned}
$$

In this study, we implemented TMLE by using the SuperLearner for the estimation of both propensity scores and outcomes [14]. The model is denoted as TMLE_SL. In addition, we implemented a model, denoted as AIPW_SL, on the basis of the AIPW method and the SuperLearner's estimation of the propensity scores and outcomes.

2.4 Simulation

The performance of the methods described in Section 2.3 for estimating ATT was compared by using Monte Carlo

Table 1. Treatment selection models.

Scenario 1	$\text{logit}(p_{i,treat}) = \alpha_{0,treat} + 0.1X_{1,i} + 0.1X_{2,i} + 0.05X_{1,i}^2 + 0.02X_{2,i}^2 + 0.02X_{1,i}X_{2,i}$
Scenario 2	$\text{logit}(p_{i,treat}) = \alpha_{0,treat} + 1.25X_{1,i} + X_{2,i} + 0.5X_{1,i}^2 + 0.5X_{2,i}^2 + 0.75X_{1,i}X_{2,i}$
Scenario 3	$\text{logit}(p_{i,treat}) = \alpha_{0,treat} + 0.1X_{1,i} + 0.1X_{2,i} + 0.05X_{1,i}^2 + 0.02X_{2,i}^2 + 0.02X_{1,i}X_{2,i} + 0.05X_{4,i} + 0.02X_{4,i}^2$

Table 2. Outcome models.

$$\textit{Setting 1} \quad Y_i = \begin{cases} Z_i + 1.5X_{1,i} + 0.75X_{3,i} + \epsilon_i, & (\textit{Scenarios 1\&2}) \\ Z_i + 1.5X_{1,i} + 0.75X_{3,i} + 5X_{4,i} + \epsilon_i, & (\textit{Scenario 3}) \end{cases}$$

$$\textit{Setting 2} \quad Y_i = \begin{cases} Z_i + 1.5X_{1,i} + 0.75X_{3,i} + 1.75X_{1,i}^2 + \epsilon_i, & (\textit{Scenarios 1\&2}) \\ Z_i + 1.5X_{1,i} + 0.75X_{3,i} + 1.75X_{1,i}^2 + 5X_{4,i} + \epsilon_i, & (\textit{Scenario 3}) \end{cases}$$

$$\textit{Setting 3} \quad Y_i = \begin{cases} Z_i + 1.5X_{1,i} + 0.75X_{3,i} + 1.5X_{1,i}Z_i + \epsilon_i, & (\textit{Scenarios 1\&2}) \\ Z_i + 1.5X_{1,i} + 0.75X_{3,i} + 1.5X_{1,i}Z_i + 5X_{4,i} + \epsilon_i, & (\textit{Scenario 3}) \end{cases}$$

simulations. To evaluate the methods in practical situations where the identifiability assumptions may be violated, we considered the following three scenarios:

- *Scenario 1:* No violation of identifiability assumptions,
- *Scenario 2:* Violation of the positivity assumption,
- *Scenario 3:* Violation of the unconfoundedness assumption.

Note that we did not consider the scenario where the consistency assumption was violated because the consistency could be ensured by proper study design and conduct.

In the simulation, the patient's treatment status was generated by the treatment selection model based on the covariates, while the patient's outcome was generated by the outcome model based on the covariates and treatment status.

In *Scenarios 1* and *2*, we assumed three independent covariates $X_1 - X_3$ that followed the standard normal distribution. X_1 was a confounder that affected both the treatment selection and the outcome, and X_2 and X_3 affected the treatment selection and the outcome, respectively. In *Scenario 3*, an additional confounder X_4 was assumed to follow the standard normal distribution and to be independent of $X_1 - X_3$.

The treatment selection models for corresponding scenarios used in our simulations are listed in Table 1. For each subject, treatment status, denoted by Z_i, was generated from a Bernoulli distribution with a probability of success of $p_{i,treat}$. We considered five different prevalence rates of treatment, i.e., expectations of $p_{i,treat}$ with respect to the distributions of covariates: 0.05, 0.10, 0.20, 0.33, and 0.50. The intercept $\alpha_{0,treat}$ in each treatment selection model was determined numerically to obtain the desired treatment prevalence.

For each scenario, we considered three settings of outcome models:

- *Setting 1:* Linear outcome model.
- *Setting 2:* Nonlinear outcome model.
- *Setting 3:* Outcome model with the interaction term between the treatment variable and the covariate.

The outcome models used in our simulations to reflect these settings are listed in Table 2, where we assumed $\epsilon_i \sim N(0, 2)$.

We performed 200 simulations for each prevalence rate of treatment under each combination of treatment selection model and outcome generation model. The steps of simulations are described below.

1. Generated N sets of covariates from independent standard normal distributions, assuming that each set was associated with one patient. The sample size N was determined by making the expected number of treated patients equal to 50, e.g., $N = 1000$ if the prevalence rate of treatment was 0.05, and $N = 500$ if the prevalence rate of treatment was 0.10.
2. For each set of the covariates associated with a patient, compute $p_{i,treat}$ by using the corresponding treatment selection model. The patient's treatment status Z_i was then generated via the Bernoulli distribution with a success probability of $p_{i,treat}$. The patient's outcome Y_i was generated by using the corresponding outcome model.
3. Estimated the ATT by using the methods described in Section 2.3. For each method, we included only covariates $X_1 - X_3$ for estimation. We used the logistic regression based on $X_1 - X_3$ to estimate the PS for PSM,

PSM_1:2, IPW, and AIPW. And we used the LR based on $X_1 - X_3$ to estimate the outcomes for AIPW.

4. Repeated the above process 200 times.
5. Evaluated the performance of each method through following metrics:

   - The bias in estimating ATTs, i.e., the average difference between the 200 estimated ATT and the true ATT. Note that the true ATTs in *Settings 1* and *2* are 1. For *Setting 3*, it is difficult to calculate the true ATT analytically and we approximated the true ATT by using a large number of simulated samples (i.e., 10^7) from the treatment models described in Table 1.

   - The empirical SD of the estimated ATTs, i.e., the SD of 200 estimated ATTs.

   - The average theoretical SD of the estimated ATTs, i.e., the average of 200 model-based SDs, each generated directly by the model in each simulation.

   - The MSE in estimating ATTs, i.e., the MSE of 200 estimated ATTs.

   - The type I error rate, i.e., the average rate of rejecting the null hypothesis of ATT $= 0$ versus the alternative hypothesis of ATT $\neq 0$ using a two-sided test at the 0.05 significance level by making $Z_i = 0$ in the outcome models described in Table 2.

Note that we set the expected number of treated subjects as 50 to reflect the practical situation that the feasible number of patients who can be enrolled in the study requiring the external control is often limited.

In this study, the nearest neighbor matching without replacement algorithm with a caliper of 0.2 was used in forming matched pairs for PSM, PSM_1:2, and MDM to improve the balance of the matching and reduce the MSEs of the estimates [5, 6]; a trimming threshold of 0.05 was used for estimated the PS for IPW and AIPW for better variance estimation [7]; a truncation threshold of $5/(\sqrt{N}\ln(N))$ was used for estimated the PS for AIPW_SL and TMLE_SL to minimize both bias and MSE of the estimates [15].

For the CEM, MDM, and PSM implemented in this study, the average theoretical SD of the ATT were estimated through the paired t-test between two groups in matched samples to account for the dependence induced by the matching [3]. For PSM_1:2, the estimator-based SD was also estimated through the paired t-test in matched samples where each treated sample is compared to the average of the two matched control samples.

For this study, all statistical analyses were conducted in R (Version 4.2.2) using the following libraries: MatchIt (Version 4.5.5), PSweight (Version 1.1.8), tmle (Version 2.0.0), and their respective dependencies.

3. RESULTS

We report results for the three different scenarios: 1) no violation of the identifiability assumptions, 2) violation of the positivity assumption, and 3) violation of the unconfoundedness assumption separately in the following subsections.

3.1 *Scenario 1*: No Violation of the Identifiability Assumptions

The distribution of the PS values for *Scenario 1* assuming no violation of the positivity or the unconfoundedness is shown in Figure S1 in the *Supplementary Material*. As shown in the figure, the distributions are bounded away from 0 and 1, and thus the positivity assumption can be considered reasonable. The bias, empirical SD (hereafter referred to as SD), average theoretical SD, MSE, and type I error rate for the 10 methods in *Scenario 1* under three different settings are shown in Figures S2-S4 in the *Supplementary Material*, respectively.

In *Setting 1* with linear effect of covariates in the true outcome model, all methods have very small biases. The matching-based methods (i.e., CEM2, CEM5, PSM, PSM_1:2, and MDM) have relatively larger SDs than the other methods for most prevalence rates of treatment, which results in greater MSEs of these methods. Among the matching-based methods, PSM_1:2 has a relatively smaller SD than PSM when the prevalence rate of treatment is less than 0.5, while CEM2 has a relatively smaller SD than CEM5. The IPW, AIPW, AIPW_SL, and TMLE_SL models in general have similar performance in terms SD and MSE. Under the null treatment effect, the type I error rates of all methods are generally controlled at 0.05.

In *Setting 2* with nonlinear effect of covariates in the true outcome model, the biases of most models increased compared with *Setting 1*. The CEM2 in general presents the largest bias while the CEM5 presents a relatively small bias, indicating that CEM model with two bins is too coarse to catch the nonlinear effect. The MDM and LR models in general yield the second and third largest biases, respectively. The AIPW_SL and TMLE_SL have the smallest biases and SDs and hence the smallest MSEs. The PSM in general has the largest SD and therefore a relatively large MSE. Under the null treatment effect, many methods in *Setting 2* have larger type I error rates than in *Setting 1* (e.g., CEM2, MDM, and LR), due to their increased biases in *Setting 2*.

In *Setting 3* with the interaction between the treatment variable and the covariate in the true outcome model, most methods yield very small biases. In terms of SD, the matching-based methods in general have relatively larger SDs than the other methods. The MSE pattern is similar to the SD pattern. Under the null treatment effect, the outcome models for *Settings 1* and *3* are the same, therefore, the findings regarding the type I error are also the same for these two settings.

3.2 *Scenario 2*: Violation of the Positivity Assumption

The distribution of the PS values for *Scenario 2* assuming violation of the positivity or the unconfoundedness is shown in Figure S1 in the *Supplementary Material*. As shown in the figure, a large proportion of subjects' PS values are distributed at or close to 0 or 1 and thus the positivity assumption is considered violated. The bias, empirical SD, average theoretical SD, MSE, and type I error rates for the 10 methods in *Scenario 2* under three different settings are shown in Figures S5–S7 in the *Supplementary Material*, respectively.

In *Setting 1* with linear effect in the true outcome models, CEM2 and IPW in general have noticeably larger biases compared to other methods. The increased bias of CEM2 in *Scenario 2* compared to *Scenario 1* is attributed to the impact of violating the positivity assumption. Under *Scenario 2*, many subjects cannot be matched exactly but are still matched when using two bins, thus causing a larger bias. The IPW also produces a relatively larger bias under *Scenario 2*. The SD pattern in *Scenario 2* differs from that in *Scenario 1* in the following ways: 1) The SDs of IPW, AIPW, and TMLE_SL have a noticeable increase, and 2) the average theoretical SD of TMLE_SL is much smaller than the empirical SD. In terms of MSE, LR and AIPW_SL are the best two performing models. Under the null treatment effect, CEM2, IPW, and TMLE_SL experience remarkable type I error inflation. According to Figure S5 in the *Supplementary Material*, the type I error inflation of the CEM2 or IPW is mainly due to the large bias, while the type I error inflation of the TMLE_SL is mainly due to its underestimation of the SD. In contrast, PSM and PSM_1:2 perform best in terms of type I error control.

In *Setting 2* with nonlinear effect in the true outcome models, in addition to the CEM2 and IPW, which are already substantially biased in *Setting 1*, LR and AIPW are also largely biased. The bias of CEM5 is relatively large when the prevalence rate of treatment is low and relatively small when the prevalence rate of treatment is high. The biases of PSM and PSM_1:2 are the smallest. The SDs of most models in *Setting 2* increase compared to *Setting 1*, especially for IPW and LR. The AIPW_SL has a relatively small bias and the smallest SD, resulting the smallest MSE. Under the null treatment effect, LR, CEM2, IPW, AIPW, and CEM5 show severe type I error inflation, mainly due to their large biases (Figure S6 in the *Supplementary Material*); TMLE_SL also shows severe type I error inflation, mainly due to its underestimation of the SD (Figure S6 in the *Supplementary Material*). AIPW_SL has relatively mild type I error inflation, while PSM, PSM_1:2, and MDM have the best type I error control.

In *Setting 3* with the interaction between the treatment variable and the covariate in the true outcome model, the IPW, LR, and matching-based methods in general have relatively large biases (Figure S7 in the *Supplementary Material*). The doubly robust methods (i.e., AIPW, AIPW_SL, and TMLE_SL) in general have relatively small biases. Among the doubly robust methods, AIPW and TMLE_SL have relatively large SDs and hence relatively large MSEs. The MSE of AIPW_SL is the smallest, due to its relatively small bias and SD. The findings for *Setting 3* regarding the type I error inflation are the same as those for *Setting 1* (Figure S7 in the *Supplementary Material*).

3.3 *Scenario 3*: Violation of the Unconfoundedness Assumption

For all methods and three settings of the outcome model, violation of the unconfoundedness assumption in *Scenario 3* leads to a certain amount of bias (or a larger bias if the method is already biased) compared to *Scenario 1* where this assumption is not violated (Figures S8–S10 in the *Supplementary Material*). The biases induced by violation of unconfoundedness assumption are all positive under our simulation settings. We conducted post-hoc analyses by varying the coefficient values of the unmeasured confounder in the three settings and found that the sign and magnitude of the induced bias are mainly determined by the coefficient value of the unmeasured confounder.

The pattern of SD is similar between *Scenario 1* and *Scenario 3*. The MSEs observed in *Scenario 3* are greater than those observed in *Scenario 1*. Under the null hypothesis, the type I error rates in *Scenario 3* generally increase compared to *Scenario 1* due to the larger biases.

4. DISCUSSION

4.1 Performance of Estimation

4.1.1 Matching-Based Methods

Among matching-based methods, our simulation results show that the PSM model in general results in a small bias when the positivity assumption is held. When the positivity assumption is violated and when there is interaction between the treatment variable and the covariate (i.e., *Scenario 2 Setting 3*), a mild bias can be induced. This is because under the violation of the positivity assumption, the integration of $E[Y_i(1)|PS(\boldsymbol{X}_i), Z_i = 1] - E[Y_i(0)|PS(\boldsymbol{X}_i), Z_i = 0]$ in equation (2.2) is carried out in the overlapped range of the PS distributions for the two groups and hence the causal effect of interest has been deviated from the ATT. Similar issues occur in the other matching-based methods that we studied. Of note, the bias of the PSM can still be small in cases with positivity violation but without interaction between the treatment variable and the covariate, e.g., *Scenario 2 Settings 1* or *2*, because in equation (2.2), $E[Y_i(1)|PS(\boldsymbol{X}_i), Z_i = 1] - E[Y_i(0)|PS(\boldsymbol{X}_i), Z_i = 0] = E[E[Y_i(1)|\boldsymbol{X}_i, Z_i = 1] - E[Y_i(0)|\boldsymbol{X}_i, Z_i = 0]|PS(\boldsymbol{X}_i)]$ is a constant, and its integration over $PS(\boldsymbol{X}_i)|Z_i = 1$ is the same as its integration in the overlapped range of the PS distribution.

The bias of PSM_1:2 was comparable to that of PSM. However, the MSE of PSM_1:2 is generally smaller than PSM due to a smaller SD.

When considering the choice of matching ratio, our simulation indicates that 1 to 2 matching provides a reasonable balance between bias reduction and precision for estimating the ATT in most of our settings with the prevalence rate of treatment ≤ 0.33. This finding aligns with the simulation results of Austin [4] for observational studies, which demonstrated that matching one or two untreated subjects to each treated subject generally provides good performance for ATT estimation. For scenarios where the control group is substantially larger than the treatment group and high-quality matches are available, using 1 to r matching with $r > 2$ may be considered to improve efficiency, though maintaining match quality is needed.

The performance of MDM was in general similar to PSM. The MDM has a relatively large bias in *Scenario 1 Setting 2* because we used Mohalanobis distance to match only first-order terms and the second-order terms were not well matched. Through simulations, we found that if we also consider the second order terms (i.e., treat X_1^2, X_2^2, and X_3^2 as three additional covariates to be matched) in the MDM, the bias of the model in *Setting 2* minimized. However, in realistic situations we may not know the functional forms of the covariates in the outcome model for matching.

Compared to PSM and MDM, which are members of the class of equal percent bias reducin (EPBR) and do not guarantee any level of imbalance reduction, CEM is a member of the class of Monotonic Imbalance Bounding (MIB) and is able to guarantee that the imbalance will not be larger than the ex ante user choice [23]. However, the selection of the number of bins can be challenging. If a large number of bins are used, the matching performance would be good, but too many samples may be discarded, resulting in inefficient inference. If a small number of bins are used, the number of matched samples may be large, but the matching performance is likely to be poor [21, 23]. As a result, substantive knowledge of the measurement scale of each variable is needed to decide how much each variable can be coarsened without losing crucial information [23]. In addition, it is difficult to perform exact matching for data with a large number variables even with very coarse bins, i.e., the curse of dimensionality. Our simulations also show that the performance of the CEM is highly dependent on the choice of the number of bins. In most of our settings, CEM2 had the largest bias and therefore the largest MSE among all models we studied. In contrast, the bias of CEM5 is generally small. However, in some settings (e.g., *Scenario 1 Setting 1*), CEM5 has a larger SD than CEM2, which leads to a larger MSE.

4.1.2 Nonmatching-Based Methods

The nonmatching methods in general have relatively smaller SDs than the matching based methods as they can utilize information from more samples.

Figure 2: Summary of MSE based on simulations under *Scenarios 1* and *2*. We evaluate models by comparing their relative performance to other models and scenario-setting combinations; therefore, identical icons do not ensure equal performance.

In many of our settings, the SD and MSE of the basic model LR are very small. However, when the nonlinear effect is present in the outcome model, the bias and the MSE of the LR model can be very large. In addition, the LR is also affected by the violation of the positivity assumption. As shown in our simulation results, the bias of the LR model under *Scenario 2 Setting 2* is much greater than its bias under *Scenario 1 Setting 2* because the model dependence issue is exaggerated through extrapolation with positivity violation [19].

For the inverse probability weighting methods, the performance of the AIPW model in our simulations is always not inferior to, if not better than, the IPW model in terms of bias and MSE, which demonstrates the advantage of the AIPW's doubly robustness in estimation accuracy. However, when the nonlinear effect is present in the outcome model (e.g., *Scenario 1 Setting 2*, *Scenario 2 Setting 2*), the AIPW model has a relatively large bias because it does not provide a good estimate of the treatment assignment or the outcome.

The bias of AIPW_SL is much smaller than the AIPW when the nonlinear effect is present and comparable to the AIPW in other settings. Attributed to the doubly robustness and SuperLearner's flexibility, the AIPW_SL always presents a small bias in all of our settings in *Scenarios 1* and *2*. In addition, the SD of AIPW_SL is also very small for all settings in *Scenarios 1* and *2*. Therefore, the MSE of AIPW_SL is always one of the smallest, if not the smallest, of all settings for *Scenarios 1* and *2*. Another doubly robust estimator with the SuperLearner, TMLE_SL, has similar bias/SD/MSE as the AIPW_SL in *Scenario 1*, but higher SD and MSE in *Scenario 2*.

A summary of MSE based on simulations under *Scenarios 1* and *2* is provided in Figure 2.

Under *Scenario 3* where an important confounder is not considered in estimating ATT, a substantial bias is intro-

Scenario	Setting	LR	PSM	PSM_1:2	MDM	CEM2	CEM5	IPW	AIPW	AIPW_SL	TMLE_SL
1	1&3	🙂	🙂	🙂	🙂	🙂	🙂	🙂	🙂	🙂	🙂
	2	🙂	🙂	🙂	🙂	😐	🙂	🙂	🙂	🙂	🙂
2	1&3	😐	🙂	🙂	🙂	☹️	😐	😐	😐	😐	☹️
	2	☹️	🙂	🙂	🙂	☹️	☹️	☹️	☹️	😐	☹️

No or relatively small 🙂　　　Medium 😐　　　Relatively large ☹️

Figure 3: Summary of type I error rates based on simulations under *Scenarios 1* and *2*. We evaluate models by comparing their relative performance to other models and scenario-setting combinations; therefore, identical icons do not ensure equal performance.

duced in all models. Therefore, the bias due to the violation of unconfoundedness cannot be mitigated by model selection. Although the unconfoundedness assumption cannot be formally examined, sensitivity analyses have been developed to assess whether a significant result is sensitive to potential unmeasured confounding. For example, a popular sensitivity analysis is based on the E-value measuring the minimum strength of association that an unmeasured confounder would need to have with both the treatment and outcome to explain away the observed causal effect [36]. It is important to note, however, that the application of the E-value is constrained by certain limitations, and its interpretation should be approached with caution [8].

4.2 Type I Error Control

In terms of type I error control, the PSM, PSM_1:2, and MDM models perform best for all settings in *Scenarios 1* and *2* because of their small biases and accurate estimation of SDs.

In certain scenario-setting combinations, LR, CEM2, CEM5, IPW, and AIPW exhibit inflated type I errors, primarily attributable to biases in their estimation.

The adequacy of type I error control in TMLE_SL is contingent on the fulfillment of the positivity assumption. Our simulations reveal that when this assumption is breached, TMLE_SL tends to underestimate the SD, resulting in severe type I error inflation. This result echoes previous findings in the literature that the use of TMLE in estimating the average treatment effect (ATE) produces type I error inflation [26]. The AIPW_SL encounters a similar issue, though its severity is less pronounced compared to the TMLE_SL.

A summary of type I error rates based on simulations under *Scenarios 1* and *2* is provided in Figure 3.

4.3 Scope and Limitations of Simulation Findings

We would like to emphasize that the primary purpose of our simulations is to identify scenarios and settings that re-

veal the limitations of the causal inference methods, rather than to provide a comprehensive evaluation under all possible conditions, which is not feasible. As a result, the comparison results for performance of estimation and type I error control may not directly generalize to other simulation settings. When considering specific methods, we recommend that trialists perform thorough evaluations tailored to their particular needs prior to implementation.

4.4 Considerations on the Design and Selecting Analytic Methods for Single-Arm Trials with External Controls

This study compared the performance of various methods for estimating the ATT in single-arm clinical trials with external RWD controls. Through simulations under different identifiability assumptions, we evaluated metrics of 10 causal inference methods, including matching-based and nonmatching-based approaches. We found that the doubly robust methods generally outperformed other approaches in terms of bias. However, no single method consistently outperformed others in terms of MSE. Additionally, the identifiability assumptions played a critical role in the models' performance, with violations of the positivity and unconfoundedness assumptions leading to potential inflation of type I errors and biases. Among the matching-based methods, PSM, PSM_1:2, and MDM showed robustness under various settings, but their performance was affected by the violation of the positivity assumption and the interaction between the treatment variable and the covariate. By contrast, nonmatching methods in general had smaller SDs than matching-based methods due to their ability to utilize more data. AIPW_SL and TMLE_SL exhibited advantages in terms of bias and MSE due to their double robustness and flexibility, but they tended to underestimate the SD and experienced type I error inflation with violated positivity assumption.

In light of these findings, it is essential to recognize the limitations of causal inference methods for single-arm trials with external controls, especially when any identifiability assumption can be violated. Violations of these assumptions can lead to biased estimates of treatment effects and an increased risk of type I errors. Therefore, it is crucial to have careful considerations of these factors in both the design and analysis phases of such trials. Accordingly, we would like to recommend the following procedures for planning the single-arm trial with an external control.

1. **Determine the causal treatment effect of interest:** To address pertinent clinical questions, it is essential to select the appropriate causal treatment effect of interest. Specifically, when seeking to understand the treatment effect solely in treated patients, attention should be directed towards ATT.

2. **Consider the consistency assumption in the design phase:** The consistency assumption requires the

treatment be sufficiently well-defined. Therefore, during the design stages, the clinical experts should scrutinize the definition of the treatment to ensure that it is adequately well-defined. Additionally, it is important to ensure that the data to be collected contain values that align with the well-defined treatment.

3. **Consider the positivity assumption in the design phase:** The positivity assumption cannot be examined at the trial design stage. However, measures should be used to make this assumption as valid as possible. If historical trial patients are selected for inclusion in the external control arm (ECA), the planned single-arm trial and historical trials should share similar inclusion/exclusion criteria and other design elements. For other sources of RWD to be included in an ECA, e.g., electronic medical records, the eligibility criteria for the single-arm and the ECA should also be made to avoid theoretical violation of the positivity assumption. An alternative and preferred way to fulfill this purpose is to use prospectively-collected RWD rather than retrospectively-collected. In addition, if possible, prior study information can be used to evaluate the positivity assumption, e.g., plotting the PS distribution based on historical studies with the same treatment.

4. **Consider the unconfoundedness assumption in the design phase:** The unconfoundedness assumption can be met if all confounders are identified and controlled. However, if an important confounding factor is not taken into account, the analysis results can be highly biased. Every effort should be made to identify all important confounders through the literature and expert knowledge. In addition, it should be ensured that the information from all important confounders is measured in the planned single-arm trial and is available in the RWD used for the ECA. Furthermore, sensitivity analyses such as the E-value or other tipping point analyses should be pre-planned during the design stage.

5. **Select appropriate analytic models for the primary and sensitivity analyses:** One of the most important requirements for the primary analysis method is to ensure the control of the type I error. After the type I error control is taken into account, the accuracy of the causal effect estimation is the next important consideration. For trials in which the positive assumption cannot be validly justified in the design phase, we suggest using methods that can largely control the type I error even if the positive assumption does not hold, e.g., the PSM or MDM model. When the expected prevalence rate of treatment is low, then 1 to r matching with $r \geq 2$ should be considered. For experiments where the positivity assumption can be validly justified, a doubly robust approach such as TMLE_SL or AIPW_SL may be recommended as the primary analysis method to obtain better estimation precision.

After the trial is conducted and the trial data are available, one may consider the following procedures to assist in analyses using the causal inference models.

1. **Check the performance of the method:** For example, balance diagnostics should always be conducted to evaluate the matching performance of the matching-based method [37].

2. **Evaluate the positivity assumption:** The plot of the PS distribution should be plotted to evaluate the positivity assumption. If the positivity assumption appears to be clearly violated, then results based on methods that are strongly influenced by the violation of this assumption, such as doubly robust methods with the SuperLearner, may not be reliable and should be interpreted with caution.

3. **Conduct sensitivity analyses for the unconfoundedness assumption:** Although the unconfoundedness assumption in general cannot be tested directly, sensitivity analyses should be performed to assess whether the primary analysis results are robust against the unmeasured confounder.

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

Supplementary materials are available online with this paper at the New England Journal of Statistics in Data Science website which includes Figures S1–S10.

FUNDING

REFERENCES

[1] ABADIE, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies* **72**(1) 1–19. https://doi.org/10.1111/0034-6527.00321. MR2116973

[2] ABDIA, Y., KULASEKERA, K., DATTA, S., BOAKYE, M. and KONG, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: a comparative study. *Biometrical Journal* **59**(5) 967–985. https://doi.org/10.1002/bimj.201600094. MR3696495

[3] AUSTIN, P. C. (2009). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics* **5**(1). https://doi.org/10.2202/1557-4679.1146. MR2504960

[4] AUSTIN, P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American journal of epidemiology* **172**(9) 1092–1097.

[5] AUSTIN, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* **10**(2) 150–161.

[6] AUSTIN, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine* **33**(6) 1057–1069. https://doi.org/10.1002/sim.6004. MR3249041

[7] AUSTIN, P. C. (2022). Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. *Statistics in Medicine* **41**(22) 4426–4443. MR4483678

[8] BLUM, M. R., TAN, Y. J. and IOANNIDIS, J. P. (2020). Use of E-values for addressing confounding in observational studies—an empirical assessment of the literature. *International Journal of Epidemiology* **49**(5) 1482–1494.

[9] CHATTON, A., LE BORGNE, F., LEYRAT, C., GILLAIZEAU, F., ROUSSEAU, C., BARBIN, L., LAPLAUD, D., LÉGER, M., GIRAUDEAU, B. and FOUCHER, Y. (2020). G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific Reports* **10**(1) 9219.

[10] CHESNAYE, N. C., STEL, V. S., TRIPEPI, G., DEKKER, F. W., FU, E. L., ZOCCALI, C. and JAGER, K. J. (2022). An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal* **15**(1) 14–20.

[11] DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)* **27**(3) 368.

[12] FANG, Y. (2020). Two basic statistical strategies of conducting causal inference in real-world studies. *Contemporary Clinical Trials* **99** 106193.

[13] FANG, Y., WANG, H. and HE, W. (2020). A statistical roadmap for journey from real-world data to real-world evidence. *Therapeutic Innovation & Regulatory Science* **54**(4) 749–757.

[14] GRUBER, S., VAN DER LAAN, M., KENNEDY, C. and GRUBER, M. S. (2006). Package 'tmle'. *Biostatistics* **2** 1.

[15] GRUBER, S., PHILLIPS, R. V., LEE, H. and VAN DER LAAN, M. J. (2022). Data-adaptive selection of the propensity score truncation level for inverse-probability–weighted and targeted maximum likelihood estimators of marginal point treatment effects. *American Journal of Epidemiology* **191**(9) 1640–1651.

[16] HECKMAN, J. J. and ROBB JR, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics* **30**(1-2) 239–267.

[17] HECKMAN, J. J., ICHIMURA, H. and TODD, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* **64**(4) 605–654. https://doi.org/10.1111/1467-937X.00044. MR1623713

[18] HERNÁN, M. and ROBINS, J. (2020) *Causal Inference: What If.* Chapman & Hall/CRC Boca Raton, FL.

[19] HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**(3) 199–236.

[20] HO, D., IMAI, K., KING, G., STUART, E. and WHITWORTH, A. (2018). Package 'MatchIt'. *Version; 2018.*

[21] IACUS, S., KING, G. and PORRO, G. (2009). CEM: Software for coarsened exact matching. *Journal of Statistical Software* **30** 1–27.

[22] IACUS, S. M., KING, G. and PORRO, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* **106**(493) 345–361. https://doi.org/10.1198/jasa.2011.tm09599. MR2816726

[23] IACUS, S. M., KING, G. and PORRO, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis* **20**(1) 1–24.

[24] IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93**(2) 126–132.

[25] KING, G. and NIELSEN, R. (2019). Why propensity scores should not be used for matching. *Political Analysis* **27**(4) 435–454.

[26] LÉGER, M., CHATTON, A., LE BORGNE, F., PIRRACCHIO, R., LASOCKI, S. and FOUCHER, Y. (2022). Causal inference in case of near-violation of positivity: comparison of methods. *Biometrical Journal* **64** 1389–1403. https://doi.org/10.1002/bimj.202000323. MR4523219

[27] LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113**(521) 390–400. https://doi.org/10.1080/01621459.2016.1260466. MR3803473

[28] MAO, H., LI, L. and GREENE, T. (2019). Propensity score weighting analysis and treatment effect discovery. *Statistical Methods in Medical Research* **28**(8) 2439–2454. https://doi.org/10.1177/0962280218781171. MR3988108

[29] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**(429) 106–121. MR1325118

[30] ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)* **45**(2) 212–218.

[31] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1) 41–55. https://doi.org/10.1093/biomet/70.1.41. MR0742974

[32] SCHULER, M. S. and ROSE, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology* **185**(1) 65–73. https://doi.org/10.2202/1557-4679.1241. MR2595112

[33] STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**(1) 1. https://doi.org/10.1214/09-STS313. MR2741812

[34] VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**(1). https://doi.org/10.2202/1557-4679.1043. MR2306500

[35] VAN DER LAAN, M. J., ROSE, S. et al. (2011) *Targeted learning: causal inference for observational and experimental data* **4**. Springer. https://doi.org/10.1007/978-1-4419-9782-1. MR2867111

[36] VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine* **167**(4) 268–274.

[37] ZHANG, Z., KIM, H. J., LONJON, G., ZHU, Y. et al. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine* **7**(1).

[38] ZHU, Y., HUBBARD, R. A., CHUBAK, J., ROY, J. and MITRA, N. (2021). Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiology and Drug Safety* **30**(11) 1471–1485.

Huan Wang. Division of Biometrics IX, OB/OTS/CDER, FDA, Silver Spring, MD, USA.
E-mail address: huan.wang@fda.hhs.gov

Fei Wu. Division of Biometrics IX, OB/OTS/CDER, FDA, Silver Spring, MD, USA.
E-mail address: fei.wu1@fda.hhs.gov

Yeh-Fong Chen. Division of Biometrics IX, OB/OTS/CDER, FDA, Silver Spring, MD, USA.
E-mail address: yehfong.chen@fda.hhs.gov