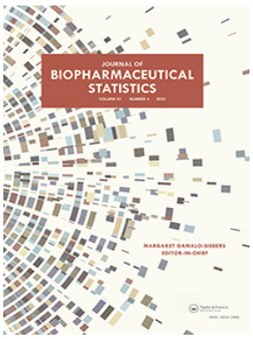


Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: www.tandfonline.com/journals/lbps20

Double machine learning methods for estimating average treatment effects: a comparative study

Xiaoqing Tan, Shu Yang, Wenyu Ye, Douglas E. Faries, Ilya Lipkovich & Zbigniew Kadziola

To cite this article: Xiaoqing Tan, Shu Yang, Wenyu Ye, Douglas E. Faries, Ilya Lipkovich & Zbigniew Kadziola (21 Apr 2025): Double machine learning methods for estimating average treatment effects: a comparative study, Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2025.2489281

To link to this article: https://doi.org/10.1080/10543406.2025.2489281

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2025.2489281

Taylor & Francis
Taylor & Francis Group

Double machine learning methods for estimating average treatment effects: a comparative study

Xiaoqing Tan[a], Shu Yang[b], Wenyu Ye[c], Douglas E. Faries[c], Ilya Lipkovich[c], and Zbigniew Kadziola[c]

[a]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA; [b]Department of Statistics, North Carolina State University, Raleigh, NC, USA; [c]Real World Analytics, Eli Lilly and Company, Indianapolis, IN, USA

ABSTRACT

Observational cohort studies are increasingly being used for comparative effectiveness research to assess the safety of therapeutics. Recently, various doubly robust methods have been proposed for average treatment effect estimation by combining the treatment model and the outcome model via different vehicles, such as matching, weighting, and regression. The key advantage of doubly robust estimators is that they require either the treatment model or the outcome model to be correctly specified to obtain a consistent estimator of average treatment effects, and therefore lead to a more accurate and often more precise inference. However, little work has been done to understand how doubly robust estimators differ due to their unique strategies of using the treatment and outcome models and how machine learning techniques can be combined to boost their performance, which we call double machine learning estimators. Here, we examine multiple popular doubly robust methods and compare their performance using different treatment and outcome modeling via extensive simulations and a real-world application. We found that incorporating machine learning with doubly robust estimators such as the targeted maximum likelihood estimator gives the best overall performance. Practical guidance on how to apply doubly robust estimators is provided.

ARTICLE HISTORY
Received 29 April 2024
Accepted 03 March 2025

KEYWORDS
Augmented inverse probability weighting; double score matching; penalized spline of propensity methods for treatment comparison; SuperLearner

1. Introduction

Randomized control trials (RCTs) are considered to be the gold standard for establishing the causal effects of interventions. They evaluate interventions among comparable groups (Hariton and Locascio 2018; Stolberg et al. 2004). However, sometimes it would be impossible to conduct RCTs due to limited resources or ethical issues. Observational studies, on the other hand, examine effects in "`real world" settings without manipulation (Rosenbaum et al. 2010). As there is no intervention, some individuals with certain characteristics may have a different probability of being exposed to treatment than others, meaning that the covariate information between treatment groups may be highly imbalanced. Therefore, it's important to adjust for covariate imbalance issues in observational studies.

There are two ways of adjustment for observational studies. The first kind is based on the treatment model, also known as the propensity score (PS) model, where the PS is defined to be the probability of being treated given covariates (Rosenbaum and Rubin 1983). The common inverse propensity treatment weighted estimator falls into this category. The idea of weighting is to create a weighted pseudo-population where treatments are "randomized". Another kind is based on the outcome model. This outcome imputation approach tries to impute the missing potential outcomes based on outcome modeling (Little and Rubin 2019). Estimators based on PS modeling require the correct treatment

CONTACT Shu Yang syang24@ncsu.edu Department of Statistics, North Carolina State University, 2311 Katharine Stinson Dr. Raleigh, NC 27695-8203, USA

model, and estimators based on outcome modeling require the correct outcome model. In practice, it's common to use linear models for PS and outcome modeling. This, however, could be problematic because linearity may be an inappropriate assumption for PS and outcomes when the response surface is nonlinear. To account for potential nonlinearity, more flexible models are needed to be considered.

Doubly robust estimators combine the above two adjustments in a fortuitous way that the causal estimator can be consistent if either the outcome model or the treatment model is correctly specified (Bang and Robins 2005). Recently, various doubly robust estimators of different kinds such as weighting, matching, and regression have been proposed in the literature to estimate the average treatment effects (ATEs) (Glynn and Quinn 2010; van der Laan and Rubin 2006; Yang and Zhang 2022; Zhou et al. 2019). Although they all use the PS and outcome mean models, they combine them differently. Also, doubly robust estimators especially the ones derived from the semiparametric efficiency theory are known to have the *rate double robustness* property (Chernozhukov et al. 2018; Kang and Schafer 2007) in the sense that they remain root-n consistent and asymptotically normal when using machine learning approaches to estimating the nuisance functions. The unaddressed question is how different doubly robust estimators perform coupled with machine learning approaches. Also, little work has been done to understand the challenges of covariate selection, overlapping of covariate distribution, and treatment effect heterogeneity for doubly robust estimators (Naimi et al. 2023).

In this paper, we compare different double machine learning methods for estimating average treatment effects, where double machine learning is referred to as the doubly robust estimator using machine learning estimators for the propensity score and outcome mean. Specifically, we review multiple popular doubly robust methods from the categories of matching, weighting, or regression, and compare their performance using different PS and outcome modeling via extensive simulations as well as a real-world application. We found that incorporating machine learning with doubly robust estimators such as the targeted maximum likelihood estimator gives the best overall performance on estimating ATEs. The main contribution of the paper is that we conduct a comprehensive evaluation of the empirical performance of estimators. Also, practical guidance in applying these estimators is provided.

The remaining paper is organized as follows: In Section 2, we discuss both doubly robust estimators and singly robust estimators in detail. Section 3 presents the extensive comparative simulations and Section 4 reports on performance of these estimators on a real-world application. Section 5 provides practical guidance and concludes the paper.

2. Methodology: singly and doubly robust estimators

2.1. Notation, assumptions, and estimand

All aforementioned methods are based on the potential outcomes framework (Neyman 1923; Rubin 1974). Let X_i be the set of observed covariates, A_i be the binary treatment indicator, and Y_i be the observed outcome for subject $i = 1,2,\ldots,n$. Let $Y_i(a)$ be the potential outcome had subject i been given a treatment assignment a, where $a = 1$ is the treatment and $a = 0$ is the control. We assume subjects are independent. The causal estimand of interest is the average treatment effect τ, which is defined as $\tau = E\{Y(1) - Y(0)\}$.

In reality, since only one of the potential outcomes is observed and another is missing, sometimes the fundamental problem of causal inference (i.e. estimating average treatment effect) is posited as a missing data problem. To estimate causal parameter τ from data with non-randomized treatment assignment, the following causal assumptions are needed.

2.1.1. Assumption 1 (stable unit treatment value)

The potential outcomes of any individual are unrelated to the treatment assignment of other individuals and there are no multiple versions of the treatment. Sometimes, this is referred to as

"consistency" in that the observed outcome is equal to the potential outcome under the actually assigned treatment, i.e., the observed outcome is given by $Y = Y(A) = AY(1) + (1 - A)Y(0)$.

2.1.2. Assumption 2 (conditional unconfoundedness or treatment ignorability)

Given covariates X, treatment assignment is independent of potential outcomes, i.e. $\{Y(0), Y(1)\} \perp A|X$.

2.1.3. Assumption 3 (overlap or common support or positivity)

All subjects are possible to receive either arm of treatment, i.e., $0 < P(A = 1|X = x) < 1 \ \forall x \in X$.

Assumption 1 ensures there is no interference among subjects and there is no multiple versions of the treatment. For observational studies, because the exposure to treatment is not controlled, treatment may be related to the way a subject might potentially respond. Assumption 2 states that it may be possible to identify all pre-treatment covariates that are predictors of treatment or outcome. If X contains all confounders, among subjects who share the same X there is no association between A and potential outcomes. A common practice, in reality, is to collect a large number of possible confounders in order to mitigate the violation of this assumption. However, including all available covariates in the analysis may introduce bias and variance of the causal effect estimator (Brookhart et al. 2006; Myers et al. 2011; Pearl 2011; Yang et al. 2020). Variable selection is hence an important procedure when estimating ATE. Assumption 3 adds a restriction on the joint distribution of treatment assignment and covariates. Overlap is an important issue in estimating treatment effects from non-randomized trials. It describes the extent to which the range of data is the same across the two treatment groups. In fact, the lack of overlap may affect all types of estimators. For matching estimators, that means it is difficult to find good matches; for weighting estimators, small overlap can result in extremely large weights; and for regression estimators, they may heavily rely on extrapolation. When Assumption 3 is violated, a common practice is to trim the sample to restrict inference to the one with sufficient overlap (Yang and Ding 2018), or to coarsen PS (Zhou et al. 2015).

Given the above stated assumptions, the ATE can be identified from observed data by conditioning on available covariates

$$\tau = E\{Y(1)\} - E\{Y(0)\} = E\{E(Y|A = 1, X) - E(Y|A = 0, X)\} \tag{1}$$

where the outer expectation is with respect to the distribution of X over the entire population.

In this section, we first review two kinds of *singly robust* estimators based on either outcome modeling or treatment modeling, respectively, in the sense that the consistency of the estimators relies on the correctness of the underlying model. Then, we review various *doubly robust* estimators that combine outcome modeling and treatment modeling in estimating ATE. Augmented inverse probability treatment weighting (AIPTW, Cao et al. 2009; Glynn and Quinn 2010; Lunceford and Davidian 2004; Robins et al. 2000) belongs to a class of weighting estimators. AIPTW is a combination of the basic inverse probability weighting estimator and a weighted average of the outcome imputation estimators. AIPTW improves the IPW estimator by fully utilizing information about both treatment and outcome. Targeted maximum likelihood estimation (TMLE, van der Laan and Rose 2011; van der Laan and Rubin 2006) is a regression estimator based on maximum likelihood estimation and includes a "targeting" step that optimizes the bias-variance tradeoff for the causal estimand. Double score matching (DSM, Yang and Zhang 2022; Zhang et al. 2021) belongs to the class of matching estimators. DSM matches on both propensity score and prognostic score. Penalized spline of propensity methods for treatment comparison (PENCOMP, Zhang and Little 2009; Zhou et al. 2019) is an example of doubly robust regression estimator. PENCOMP estimates causal effects by imputing missing potential outcomes with flexible spline models using multiple imputations.

2.2. Estimators for ATE based on outcome modeling

A traditional way to adjust for covariate imbalance in observational studies is via the formulation of a regression model for the outcome Y on A and X. That is, we can estimate the regression E(Y |A,X) by modeling on the observed data. Given the stated assumptions Section 2.1, the ATE can be identified by Equation 1.

An example of a regression imputation estimator can be obtained by fitting a linear regression of Y given A and all X, which is given by

$$E(Y|A,X) = \alpha_0 + \alpha_A A + X^T \alpha_X. \tag{2}$$

Suppose this is indeed the true regression model, by Equation 1, $\tau = \alpha_0 + \alpha_A \cdot 1 + X^T \alpha_X - (\alpha_0 + \alpha_A \cdot 0 + X^T \alpha_X) = \alpha_A$. The ATE can be obtained directly from the coefficient for A, i.e. α_A in the regression model. If the true regression is specified, this estimator is consistent of τ. Hence, this regression imputation estimator is a singly robust estimator in the sense that it is consistent when the outcome model is correctly specified.

However, model (2) assumes a constant treatment effect and could be severely biased in the case of heterogeneous treatment effects. In practice, treatment effects may vary across subjects. The regression imputation estimator is usually obtained by modeling outcome separately within each treatment arm rather than by using a single model (2). Note that the regression above can be made more general as a general parametric model, since Y can be of any type. The missing potential outcomes are then imputed by their predictions from the corresponding posited models. The regression imputation estimator is given by the difference in the averages of potential outcomes. This can help to address heterogeneity in treatment effects; however, the issue of model misspecification still exists. Also, in the case of a near violation of Assumption 3, the outcome model-based approaches rely on extrapolation.

2.3. Estimators for ATE based on treatment modeling

Another class of ATE estimators for covariate adjustment relies on the treatment model, or the propensity score model. The propensity score is defined as the probability of treatment given covariates, i.e., $e(X) = E(A|X) = P(A = 1|X)$.

Under the assumptions defined in Section 2.1, given the propensity score, the potential outcomes and treatment assignment are independent (Rosenbaum and Rubin 1983), i.e., $\{Y(0), Y(1)\} \perp A \,|e(X)$. Traditionally, the estimation of propensity is by using a logistic regression where $e(X,\beta) = \text{logit}^{-1}\{\exp(\beta_0 + X^T \beta_X)\}$.

Consider the inverse of the propensity score as a weight for the outcome, under the assumptions stated in Section 2.1 and the true propensity score

$$\begin{aligned} E\left\{\frac{(ZY)}{e(X)}\right\} &= E\left\{\frac{(ZY(1))}{e(X)}\right\} = E\left[E\left\{\frac{(ZY(1))}{e(X)}|Y(1),X\right\}\right] = E\left\{\left(\frac{Y(1)}{e(X)}\right)E(Z|Y(1),X)\right\} \\ &= E\left\{\left(\frac{Y(1)}{e(X)}\right)E(Z|X)\right\} = E\left\{\left(\frac{Y(1)}{e(X)}\right)e(X)\right\} = E\{Y(1)\} \end{aligned}$$

Similarly, $E\{(1-Z)Y/e(X)\} = E\{Y(0)\}$.

A well-known common estimator based on propensity score is Inverse Probability Treatment Weighting (IPTW) (Lunceford and Davidian 2004). Specifically, IPTW estimates τ by the difference of inverse probability of treatment weighted averages, which is given by

$$\hat{\tau}_{IPTW} = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{(1-\hat{e}(X_i))}.$$

The inverse weights create a pseudo-population where there is no confounding so the weighted averages can reflect averages in the target population. If the true model for the propensity score model is specified, $\hat{\tau}_{IPTW}$ is a consistent estimator of τ. Hence, IPTW is a singly robust estimator in the sense that it is consistent when the treatment model is correctly specified. A major drawback of the IPTW estimator is that IPTW is highly unstable due to weighting by the inverse of the propensity score. If any propensity score is close to zero or one, the IPTW estimator may be extreme.

Variable selection is an important consideration when constructing propensity scores. Including predictors of treatment but not outcome, i.e., instrument variables, in the treatment model or the outcome model may amplify bias and variance of the causal estimator (Myers et al. 2011; Pearl 2011). Including outcome predictors, on the other hand, could boost efficiency (Brookhart et al. 2006; Tan et al. 2022; Yang et al. 2020). Therefore, variable selection is needed before the estimation of treatment effects to remove variables not related to outcomes. Besides, weighting estimators are inferior in the case of extreme propensity scores (Kang and Schafer 2007). Poor overlap in propensity score distributions can result in extremely large weights, leading to an unstable estimator with a large variance. Furthermore, model misspecification of the propensity score would also lead to a biased causal estimate.

2.4. Augmented inverse probability treatment weighted (AIPTW)

AIPTW estimator is a weighting-based estimator that improves IPTW by fully utilizing information about both the treatment assignment and the outcome (Glynn and Quinn 2010). It is a combination of IPTW estimator and a weighted average of the outcome imputation estimators. Specifically, AIPTW is given by

$$\hat{\tau}_{AIPTW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left\{\frac{A_i Y_i}{\hat{e}(X_i)} - \frac{(1-A_i)Y_i}{1-\hat{e}(X_i)}\right\} - \frac{A_i - \hat{e}(X_i)}{\hat{e}(X_i)\{1-\hat{e}(X_i)\}}[\{1-\hat{e}(X_i)\}\hat{m}_1(X_i) + \hat{e}(X_i)\hat{m}_0(X_i)]\right\} \tag{3}$$

where $m_1(X)$ is a postulated model for $E(Y|A=1,X)$ and $m_0(X)$ is a postulated model for $E(Y|A=0,X)$. The first line of equation (3) is the same as $\hat{\tau}_{IPTW}$ and the rest adjusts this estimator by a weighted average of the two outcome imputation estimators. Rearranging terms in equation (3), $\hat{\tau}_{AIPTW}$ can be given by

$$\hat{\tau}_{AIPTW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i Y_i}{\hat{e}(X_i)} - \frac{A_i - \hat{e}(X_i)}{\hat{e}(X_i)}\hat{m}_1(X_i)\right\} - (1/n)\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\hat{e}(X_i)} + \frac{A_i - \hat{e}(X_i)}{(1-\hat{e}(X_i)}\hat{m}_0(X_i)$$

AIPTW is a doubly robust estimator in that as long as either the outcome model is correct or the propensity score model is correct, $\hat{\tau}_{AIPTW}$ is a consistent estimator for τ (Glynn and Quinn 2010). Also, it enjoys good large-sample theoretical properties that it can be shown to be asymptotically normally distributed via derivation through the theory of M-estimation. Bootstrap can also be used to obtain the standard error estimates (Imbens 2004). These standard error estimators tend to be reasonable unless the estimated propensity scores are very extreme as weighting estimators are inferior in the case of extreme propensity scores (Kang and Schafer 2007). In these scenarios, AIPTW is less robust to data sparsity and near violation of the positivity assumption (Glynn and Quinn 2010).

Recently, machine learning has gained popularity in the field of causal inference (Peters et al. 2017; Prosperi et al. 2020; Tan 2023; Tan et al. 2022, 2022). Chernozhukov et al. (2018) shows that the regression imputation and IPTW estimators using machine learning nuisance function estimators tend to have large finite sample biases. AIPTW, derived from the semiparametric efficiency theory, on the other hand, enjoys the *rate double robustness* when combined with machine learning (Chernozhukov et al. 2018; Kang and Schafer 2007). That is,

these doubly robust estimators remain root-n consistent and asymptotically normal when using machine learning approaches to estimating the nuisance functions. Incorporating AIPTW with machine learning could help to mitigate the impact of regularization bias and overfitting on causal estimate (Chernozhukov et al. 2018).

2.5. Targeted maximum likelihood estimation (TMLE)

TMLE, introduced by van der Laan and Rubin (2006), is a maximum likelihood-based estimator that incorporates a "targeting" step that optimizes the bias-variance tradeoff for the targeted estimator, i.e., ATE. Specifically, TMLE obtains initial outcome estimates via outcome modeling and propensity scores via treatment modeling, respectively. These initial outcome estimates are then updated to reduce the bias of confounding, which generates the so-call "targeted" predicted outcome values.

The estimation steps of TMLE are given as follows: First, initial outcome estimates are constructed by $\hat{Y}_1 = \hat{E}(Y|A=1,X)$ and $\hat{Y}_0 = \hat{E}(Y|A=0,X)$, respectively, and the propensity scores $\hat{e}(X)$ are estimated through treatment modeling. Then, the targeted steps begin by first calculating the inverse propensity H_a for each subject,

$$H_1(A=1,X) = \{\hat{e}(X)\}^{-1} \text{ and } H_0(A=0,X) = -\{1-\hat{e}(X)\}^{-1}.$$

This is similar in form to inverse probability weights. Then, for the treatment arm and the control arm, separately, the observed outcome Y is regressed on those estimated inverse propensity with fixed intercepts. Take a binary outcome as an example. A logistic transform can be applied with a binary outcome Y where $logit\{E(\mathrm{Y}|\mathrm{A}=1,\mathrm{X}) = logit(\hat{\mathrm{Y}}_1) + \varepsilon_1 H_1$ and $logit\{E(\mathrm{Y}|\mathrm{A}=0,\mathrm{X}) = logit(\hat{\mathrm{Y}}_0) + \varepsilon_0 H_0$.

In this way, we are able to generate updated, or so-called "targeted" estimates of the set of potential outcomes, incorporating information from the treatment mechanism in order to reduce the bias. The predicted outcomes are then updated to be

$$logit\left(\hat{\mathrm{Y}}_1^*\right) = logit\left(\hat{\mathrm{Y}}_1\right) + \hat{\varepsilon}_1 H_1 \text{ and } logit\left(\mathrm{Y}_0^*\right) = logit\left(\hat{\mathrm{Y}}_0\right) + \hat{\varepsilon}_0 H_0$$

The final estimates is given by calculating ATE as mean difference in targeted predicted outcome pairs across individuals

$$\hat{\tau}_{TMLE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{1i}^* - \hat{Y}_{0i}^*\right) \tag{4}$$

The variance of the TMLE estimator is obtained based on the efficient influence curve (Díaz and van der Laan 2011; Porter et al. 2011; van der Laan and Rubin 2006; van der Laan et al. 2007). In general, TMLE and AIPTW are both efficient and have the minimum asymptotic variance under the large-sample theory. However, it has been shown that under finite sample size or challenging scenarios such as misspecified models, and nearly violated positivity, TMLE may still provide causal estimates in the range of ATE since $\hat{Y}_a^\star$ in Equation 4 is range-preserving while AIPTW may not (Pirracchio et al. 2015; Porter et al. 2011; van der Laan and Rose 2011). The estimation of TMLE is usually coupled with SuperLearner (Díaz and van der Laan 2011; van der Laan et al. 2007) for $\hat{Y}_a$ and $\hat{e}(X)$, which is an ensemble of multiple statistical and machine learning models. It learns an optimal weighted average of these models by giving higher weights to more accurate models and has been proven to have high accuracy (Díaz and van der Laan 2011; van der Laan et al. 2007). The performance of TMLE continues to boost with the help of SuperLearner. Note that hybrid estimators have been proposed to resemble TMLE and AIPTW that use coarsened propensity score estimates instead of model-

based ones (Zhou et al. 2015). They may have better performance in case of severe model misspecification. However, the choice of coarsening mechanism and the coarsening parameters may introduce extra challenges (Zhou et al. 2015), and thus these hybrid estimators are not included for comparison in later simulation studies.

2.6. Double score matching (DSM)

Matching methods are powerful because they can be used to re-create a randomized trial that is hidden under the observational study. DSM is a matching-based estimator that uses the double balancing properties of propensity score $e(X)$ and prognostic score $\Phi(X)$, the latter obtained via outcome modeling, before the matching is conducted (Yang and Zhang 2022). The prognostic score is defined as a balancing score where $\{Y(0), Y(1)\} \perp A|\Phi(X)$ (Hansen 2008). The combination of $e(X)$ and $\Phi(X)$, the double score, is also shown to be a balancing score (Antonelli et al. 2018). That is, $\{Y(0), Y(1)\} \perp A|\{e(X), \Phi(X)\}$.

DSM estimator enjoys the double robustness property in that this result holds even if only one score is correctly specified. For unit i, the potential outcome under A_i is the observed outcome Y_i. The potential outcome under $1 - A_i$ is not observed but can be imputed by the observed outcomes of the nearest M units with $1 - A_i$. Denote the augmented score $S = \{e(X), \Phi(X)^T\}^T$ as the matching variable, $J_{S,i}$ as the index set for these matched subjects for subject i, and $K_{S,i} = \Sigma_{j=1}^{n} I\left(i \in J_{S,i}\right)$ as the number of times subject i is used as a match. The matched set is constructed with distance metric such as Mahalanobis distance on S that combines propensity score and prognostic score. The initial DSM estimator of τ is

$$\hat{\tau}_{DSM}^{(0)} = \frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)\left(1 + M^{-1}K_{S,i}\right)Y_i$$

A de-biasing DSM estimator $\hat{\tau}_{DSM}$ suggested by Yang and Zhang (2022) further corrects bias by the method of sieves. Correctly the bias may help to improve finite sample performance in practice although this bias is asymptotically negligible (Yang and Zhang 2022). A wild bootstrap procedure is used to obtain the confidence interval based on Otsu and Rai (2017) for matching estimators (Yang and Zhang 2022).

Matching methods tend to be more stable tools when the propensity score is extreme (Stuart 2010). Matching estimators are robust to model misspecifications if the misspecified model belongs to the class of covariate scores (Waernbaum 2012). DSM is robust against model misspecification of either the propensity score model or the prognostic score model (Antonelli et al. 2018; Yang and Zhang 2022). Specifically, DSM provides multiple protections to model misspecification by positing multiple candidate models for both propensity and prognostic scores. This helps DSM to achieve near nominal coverage even under model misspecification (Yang and Zhang 2022). Furthermore, DSM can serve as a dimensional reduction tool in high-dimensional confounding. However, adding too many covariates could result in potential bias as matching estimators may not work well on high-dimensional covariates (Abadie and Imbens 2006). It is also pointed out that the number of posited models and their functional forms affect the efficiency of DSM in a complex way, resulting in an unstable performance if there is a large number of working models (Yang and Zhang 2022; Zhao and Yang 2021). Hence, variable selection in matching estimators is needed to help identify outcome predictors for better efficiency and remove bias in estimating the ATE. Zhang et al. (2021) investigate the performance of DSM under different strategies of variable selection, using a caliper, and matching with or without replacement, providing the best practice. Also, as the success of matching depends on the functional forms of posited models, flexible machine learning methods can be adopted in the modeling of propensity scores and prognostic scores before matching.

2.7. Penalized spline of propensity methods for treatment comparison (PENCOMP)

PENCOMP is a type of regression- and multiple imputation-based approach (Zhou et al. 2019). It builds on the method of the penalized spline of propensity prediction previously used in missing data problems (Little and An 2004; Zhang and Little 2009). Specifically, PENCOMP obtains propensity score $e(X)$ via treatment modeling and uses spline-based regressions with propensity score included for outcome modeling. Under the assumptions in Section 2.1, PENCOMP has a double robustness property for estimating ATE.

PENCOMP uses Rubin's rules for combining multiply imputed datasets. The method first generates a bootstrap sample b from the original data stratified on the treatment group. The propensity score $e(X)$ is then estimated. Then, the potential outcomes for the treatments not assigned to subjects are predicted with regression models that include splines on the logit of the propensity to be assigned that treatment, plus other outcome predictors. For each treatment group, each regression model is fitted separately. Specifically, the regression model fitted for subjects with $A_i = a$ is given by

$$E\{Y_i(a)|X_i,A_i = a\} = s\{\hat{e}(X_i)\} + g\{\hat{e}(X_i),X_i\},\ i \in \{i : A_i = a\}$$

where $s\{\hat{e}(X)\}$ is a penalized spline with pre-specified knots, and $g\{\hat{e}(X),X\}$ is a parametric function of outcome predictors as well as the estimated propensity. The missing potential outcome of a subject is then imputed based on the predictive distribution of $E\{Y\ (a)|X,A\}$. The bootstrap estimate $\hat{\tau}_{PENCOMP}^{(b)}$ is the difference in the treatment means based on the observed outcome and the imputed values of Y. The above procedure is repeated multiple (B) times. The final ATE estimate. The confidence interval of $\tau_{PENCOMP} = B^{-1}\Sigma_{b=1}^{B}\hat{\tau}_{PENCOMP}^{(b)}$ is generated from this procedure.

According to Zhou et al. (2019), PENCOMP achieves comparable performance with AIPTW in terms of bias, RMSE, and coverage under settings of low confounding and correctly specified model with linear settings. Also, PENCOMP has some advantages in nonlinear settings compared to AIPTW (Zhou et al. 2019). However, in the case of model misspecification, there exists a severe overcoverage with wider confidence interval for PENCOMP compared to AIPTW even under a linear setting (Zhou et al. 2019). It was also pointed out by Kang and Schafer (2007) that if regression models are misspecified, doubly robust methods could suffer from larger bias compared to singly robust methods. Besides, the choice of the splines and knots can be challenging in practice.

3. Simulation studies

In this section, we design Monte Carlo simulations to compare each doubly robust method under different scenarios mimicking real-world data where there is high nonlinearity in the relationship between covariates and treatment, and the relationship between covariates and outcome. Specifically, we consider settings with complex data generative models with multivariate covariates. We also consider cases of different degrees of separation of the propensity score distributions where the propensity scores may be close to zero or one. In Table 1, we provide open-source software or code that implements the surveyed DR estimators. The code of our simulation studies can be found on GitHub (https://github.com/ellenxtan/RealWorld-DoublyRobustML).

Table 1. Open-source software or code that implements the surveyed DR estimators.

DR estimators	Open source code or package
AIPTW	`AIPW` Zhong et al. (2021)
TMLE	`tmle` Gruber and van der Laan (2012)
DSM	`dsmatch` Yang et al. (2020)
PENCOMP	`PENCOMP` Zhou et al. (2019)

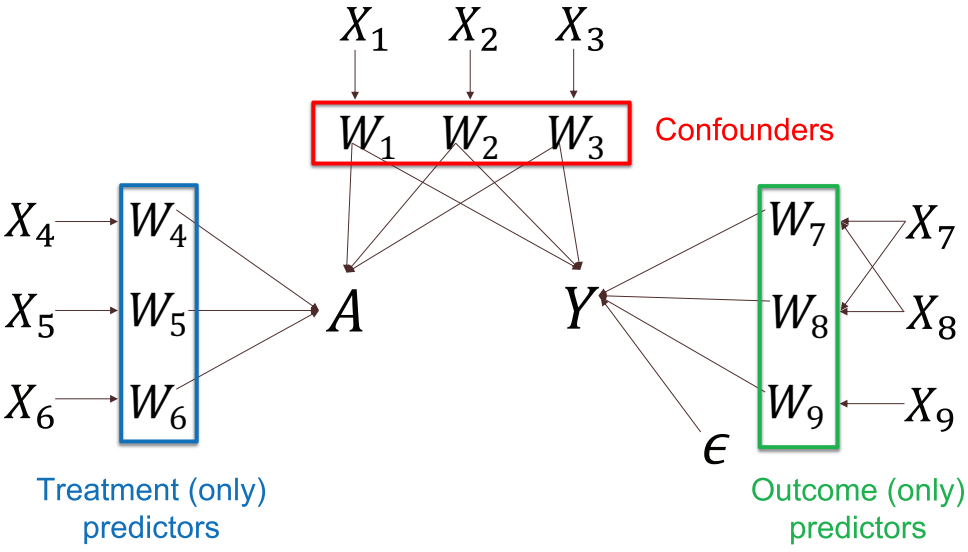


Figure 1. Illustration of variables involved in simulation studies.

3.1. Data generating process

We design the data generating procedure following the work of Leacy and Stuart (2014), but with the nonlinearity of treatment and outcome surfaces considered. The sample size is set to be $n = 2000$ throughout. The relationship of variables generated in simulation is illustrated in Figure 1. First of all, we generate $m = 9$ independent and identically distributed standard normal variables $X \in R^9 \sim N(0,1)$. Let $W \in R^9$ be a nonlinear transformation of X where W_1, W_2, W_3 are confounders (i.e., predictors of both treatment assignment and outcome), W_4, W_5, W_6 are treatment (only) predictors or instrumental variables, W_7, W_8, W_9 are outcome (only) predictors. Specifically, we let

$$W_1 = exp(X_1/2), W_2 = exp(X_2/3), W_3 = X_3^2, W_4 = X_4^2, W_5 = X_5, W_6 = X_6, W_7 = X_7 + X_8, W_8 = X_7^2 + X_8^2, W_9 = X_9^3$$

All W are standardized to have mean zero and variance 1.

We generate the binary treatment indicator A following Bernoulli$\{e(X)\}$ where $e(X)$ is the propensity score model. The propensity score model is designed to be a linear combination of confounders and treatment predictors in terms of W. To assess the property of different degrees on the separation of propensity score distributions, we consider two settings for the PS overlap to test the performance of estimators. A large overlap is an ideal case, which means there is a reasonable amount of common support for the treated subjects and the control subjects. The case of a small overlap, on the other hand, indicates a majority of the treated subjects may fail to find any suitable control in their neighborhoods. This could be common in scenarios such as rare diseases in practice. Specifically, we design the large overlap case to be $logit^{-1}\{e(X)\} = (-3 - W_1 + 2W_2 - 3W_3 + 3W_4 + 2W_5 + W_6)/15$. The distribution of the resulting propensity score is given in Figure 2. We design the case of a small overlap to be $logit-^{1}\{e(X)\} = (-8W_1 + 1.5W_2 + 0.5W_3 - 0.5W_4 + 2.5W_5 - 0.5W_6)/5$. The distribution of the resulting propensity score is given in Figure 3.

We generate a continuous outcome, which is a linear combination of confounders and outcome predictors in terms of W. Specifically, we let $Y(0) = -2 + 1.5W_1 - 2W_2 + 1.5W_3 + 2.5W_7 - W_8 + W_9 + \epsilon$ where $\epsilon \sim N(0,1)$. We assume that only X's are included in the candidate set, therefore nonlinearity in the outcome and treatment models induced by W's should be explicitly modeled by the analyst. This mimics that in practice there exists a potential nonlinearity between covariates and outcome and this nonlinearity is needed to be considered when fitting models. We consider both a homogeneous

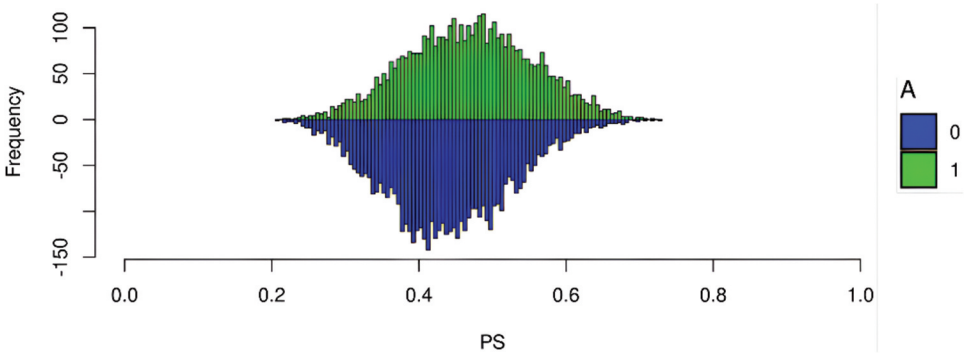


Figure 2. Distribution of the propensity score under the large overlap case.

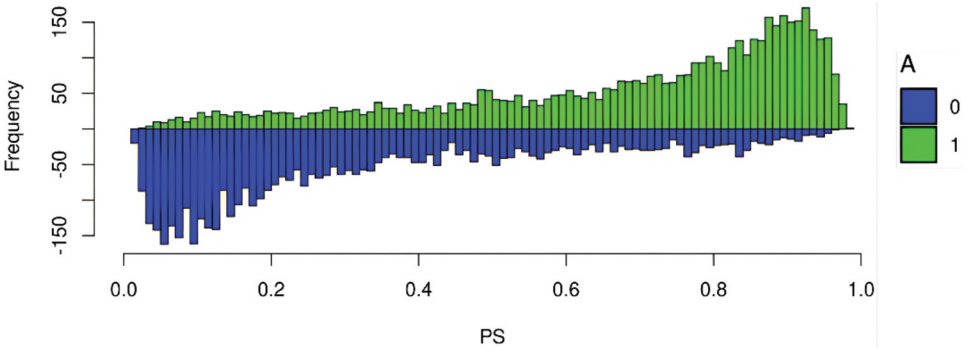


Figure 3. Distribution of the propensity score under the small overlap case.

treatment effect setting and a heterogeneous treatment effect setting. For the homogeneous setting, we consider a constant treatment effect. We use $\tau = 0$ in our simulation. Specifically, $Y(1) = Y(0) + \tau$. For the heterogeneous setting, we allow the treatment effect to vary by some covariates, and here W_1 and W_3. Specifically, $Y(1) = Y(0) + \tau + 5W_1 + 3W_3 + 2W_1W_3$. We try $\tau = 0$. Figure 4 shows how W_1 and W_3 affect the treatment effect. The observed outcome Y is given by $Y = Y(1)A + Y(0)(1 - A)$.

To summarize, there are four scenarios in total, which are

(1) homogeneous treatment effect with a large PS overlap (Homo TE + large overlap)
(2) homogeneous treatment effect with a small PS overlap (Homo TE + small overlap)
(3) heterogeneous treatment effect with a large PS overlap (Hetero TE + large overlap)
(4) heterogeneous treatment effect with a small PS overlap (Hetero TE + small overlap)

3.2. Evaluation metrics

We consider the following metrics averaged over observations and simulation replications.

- Bias: $bias = R^{-1} \sum_{r=1}^{R} \left\{ \hat{\tau}^{(r)} - \tau \right\}$
- MSE: $mse = R^{-1} \sum_{r=1}^{R} \left\{ \hat{\tau}^{(r)} - \tau \right\}^2$
- Confidence interval coverage: $coverage = R^{-1} \sum_{r=1}^{R} 1\left\{ \tau \in \left(\hat{\tau}^{(r)}_{L,0.05}, \hat{\tau}^{(r)}_{U,0.05} \right) \right\}$

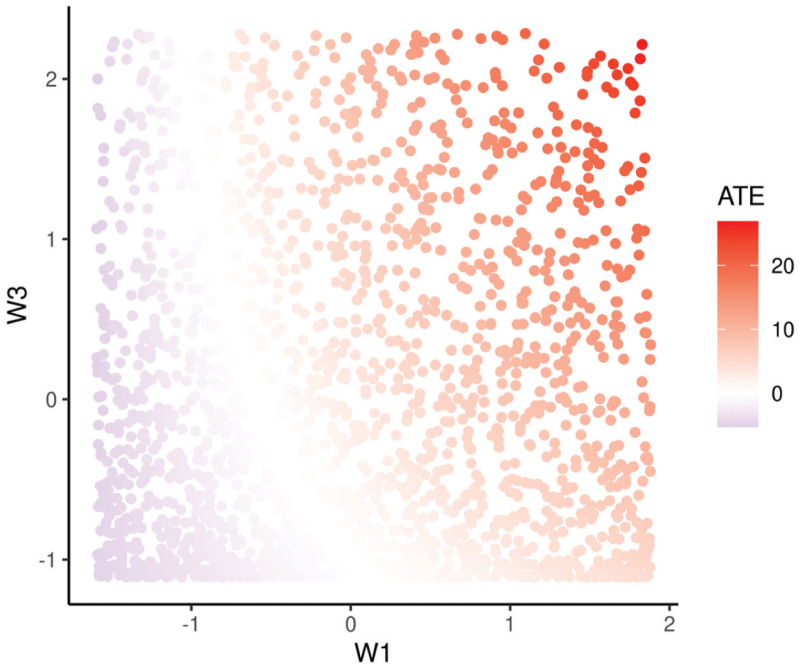


Figure 4. Treatment effect under heterogeneous treatment effect setting.

- Confidence interval width: $width = R^{-1} \sum_{r=1}^{R} \left\{ \hat{\tau}_{U,0.05}^{(r)} - \hat{\tau}_{L,0.05}^{(r)} \right\}$
- Type I error: $\alpha = 1 - R^{-1} \sum_{r=1}^{R} 1\left\{ 0 \in \left(\hat{\tau}_{L,0.05}^{(r)}, \hat{\tau}_{U,0.05}^{(r)} \right) \right\}$
- Variance ratio: $\text{var.}ratio = R^{-1} \sum_{r=1}^{R} \left[\text{var}_m \left\{ \hat{\tau}^{(r)} \right\} \right] \{ \text{var}_b(\hat{\tau}) \}^{-1}$ where $var_m(\hat{\tau}^{(r)})$ is squared estimated standard error $\hat{\tau}^{(r)}$ for the r-th replication and $var_b(\hat{\tau})$ is variance of $\hat{\tau}$ over R replications. Variance ratio measures the ratio between the mean variance of an estimator over R replicates and the variance of estimates from R replicates. It evaluates the performance of model-based standard errors of $\hat{\tau}$ by comparing them with simulation variance reflecting the true variability of estimated τ.

3.3. Analysis steps and compared estimators

For analysis steps, we generate simulated data of $n = 2000$ subjects ($n = 1000$ for treatment and control, respectively). We apply Lasso (Tibshirani 1996), a variable selection technique before the estimation in order to remove variables that are not related to outcomes. The outcome predictors are chosen by using all X's as covariates and the observed outcomes as response. Five-fold cross-validation is used to select the best tuning parameter in Lasso, with the cross-validation deviance within 1 standard error of the minimum, as recommended by Zhang et al. (2021). Outcome predictors are obtained for each treatment arm, respectively. We then use GLM, GAM, or SuperLearner separately to model PS and/or outcome before estimating the final ATE. The candidate learners in the SuperLearner library are: linear regression, stepwise regression, GAM, and Bayesian Additive Regression Trees (Chipman et al. 2010). Doubly robust estimators for comparison are AIPTW, TMLE, DSM, and PENCOMP. We also include for comparison the singly robust estimator IPTW, and a regression imputation estimator, denoted as IMP, which fits a twin outcome

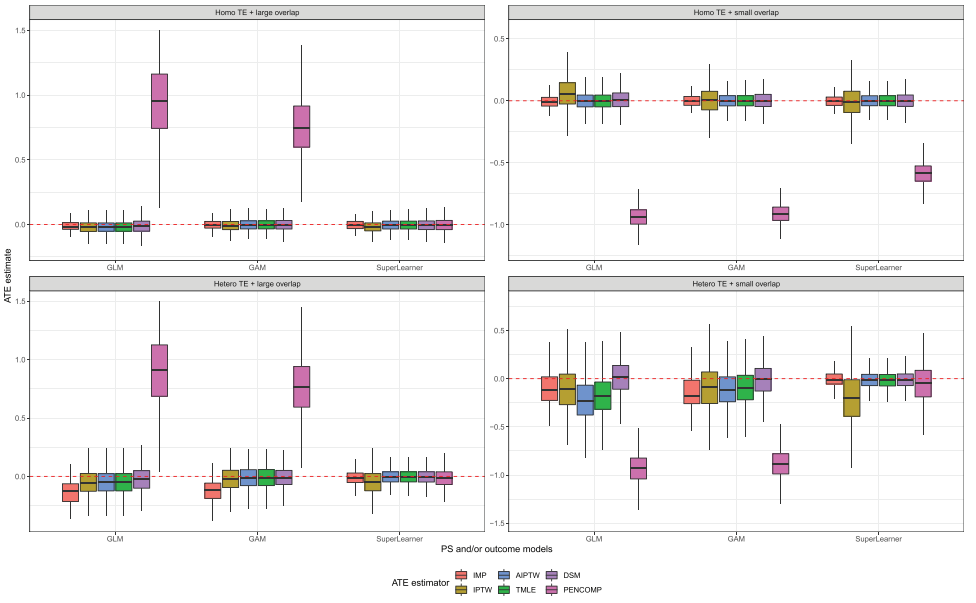
model on covariates separated by the treatment group and the control group and then imputes the missing potential outcomes with the posited models. Note that IPTW involves fitting the PS model only. We generate $R = 1000$ replications of simulated data sets. For AIPTW and TMLE, bootstrap CIs are also computed. Each replication uses $B = 500$ for the bootstrap CI.

3.4. Simulation results

Figure 5 shows the performance of various causal estimators under the four simulation scenarios. A summary of the performance of the estimators is given in Figure 6. In general, doubly robust estimators outperform singly robust estimators. IMP tends to have robust performance by chance under homogeneous settings but may fail under heterogeneous settings with GLM and GAM because of extrapolation and model misspecification. With SuperLearner, IMP greatly reduces bias and variance. IPTW has improved performance with GAM in terms of bias and MSE. However, with GLM and SuperLearner, IPTW may suffer from large bias and MSE. IPTW may achieve a small bias under homogeneous treatment effects but fails under settings of heterogeneous treatment effects. Compared to overlap between propensity score distribution, treatment effect heterogeneity has a larger effect on the performance of doubly robust estimators. Using SuperLearner for treatment and outcome modeling, doubly robust estimators achieve the smallest bias and MSE. Using GLM for treatment and outcome modeling could suffer from huge bias because the relationship between covariates and the outcome is nonlinear. Using GAM would help improve the bias and MSE (bias in particular), but some nonlinearity may still be hard to capture. Under the relatively easy homogeneous treatment effect setting, doubly robust estimators including AIPTW, TMLE, and DSM achieve a nominal confidence interval. Under challenging settings such as the heterogeneous treatment effect, all doubly robust estimators suffer from overcoverage. Overall, TMLE and AIPTW enjoy the most favorable performance with SuperLearner. They have minimal bias and MSE, especially with SuperLearner. Under the relatively simple homogeneous treatment effect setting, TMLE and AIPTW achieve nominal coverage, and with SuperLearner, these two estimators even achieve the smallest CI width and control type I error under 5%. Under the challenging settings, SuperLearner could inflate the variance ratio and show an overcoverage issue. With GAM, TMLE, and AIPTW may be able to achieve a nominal confidence interval by using bootstrap (see in the setting of heterogeneous treatment effects with a small overlap). However, in terms of model misspecification, TMLE is more robust than AIPTW with GLM and GAM. TMLE tends to have a smaller bias and MSE. DSM is more robust in terms of bias in the case of model misspecification compared to other doubly robust estimators (typically revealed in settings of heterogeneous treatment effects). DSM with GAM may outperform DSM with SuperLearner in terms of bias and variance ratio (see in the setting of heterogeneous treatment effects with a large overlap). PENCOMP suffers from severe bias and MSE with GLM or GAM. PENCOMP has improved performance with SuperLearner in terms of bias and MSE. However, PENCOMP suffers from severe overcoverage and large type I error in nearly all settings. This may be related to previous studies where in the presence of model misspecification, PENCOMP tends to exhibit substantial overcoverage, resulting in wider confidence intervals compared to AIPTW, even in a linear setting (Zhou et al. 2019). Additionally, Kang and Schafer (2007) highlight that when regression models are misspecified, doubly robust methods may experience greater bias than singly robust methods.

4. A real-world application

We apply different causal estimators to a real-world application, the Reflections study (REFL), which is a study of real-world examination of fibromyalgia for longitudinal evaluation of costs and treatments (Robinson et al. 2012). We focus the analysis on opioid treatment arm (OPI cohort), and non-narcotic opioid-like treatment arm (TRA cohort). There are 544 patients in total. The outcome of interest is the change from baseline to LOCF in the total score of the Fibromyalgia



(a)

(b)

Figure 5. Performance of various causal estimators under the four simulation scenarios. (a) Box plots of the estimators. The red dotted lines indicate the ground truth ATE. (b) Performance of the estimators in terms of CI coverage and width, type I error, and variance ratio. The red dotted lines indicate the ideal value or threshold of corresponding metrics. Methods with a "_b" indicate a bootstrap CI coverage.

		1) Homo TE + large overlap			2) Homo TE + small overlap			3) Hetero TE + large overlap			4) Hetero TE + small overlap		
		Bias	MSE	Coverage / Type I error	Bias	Variance	Coverage / Type I error	Bias	MSE	Coverage / Type I error	Bias	Variance	Coverage / Type I error
Singly robust estimators	IMP	Nearly unbiased with GLM/ GAM/SL	Small MSE with GLM/ GAM/SL	Undercoverage and large type I error with GLM/GAM/SL	Nearly unbiased with GLM/GAM /SL	Small MSE with GLM/GAM /SL	Overcoverage and type I error < 5% with GLM/GAM/SL	Huge bias with GLM/ GAM. Nearly unbiased with SL	Small MSE with SL. Large MSE with GLM/GAM	Overcoverage with SL, undercoverage with GLM/GAM	Huge bias with GLM/GAM. Nearly unbiased with SL	Small MSE with SL. Large MSE with GLM/GAM	Overcoverage with SL, undercoverage with GLM/GAM
	IPTW			Nominal coverage and type I error ~5% with GAM	Large bias with GLM. Nearly unbiased with GAM/SL	Large MSE with GLM/GAM /SL	Nominal coverage and type I error around 5% with GAM	Huge bias with GLM/SL. Smaller bias with GAM	Larger MSE than other estimators	Overcoverage. Type I error < 5%	Huge bias with GLM/GAM/SL	Larger MSE than other estimators	Undercoverag e. Type I error > 5%
Doubly robust estimators	AIPTW	Nearly unbiased with GLM/ GAM/SL	Small MSE with GLM/ GAM/SL	Nominal coverage and type I error ~5% with GAM/SL	Nearly unbiased with GLM/ GAM/SL	Small MSE with GLM/ GAM/SL	Nominal coverage and type I error ~5% using bootstrap CIs with GAM/SL	Nearly unbiased with GAM/SL. Large bias with GLM	Small MSE with SL. Large MSE with GLM/GAM	Overcoverage and type I error < 5% with GLM/GAM/SL	Nearly unbiased with SL. Larger bias than TMLE when with GLM/GAM	Small MSE with SL. Large MSE with GLM/GAM	Overcoverage with GAM/SL, undercoverage with GLM
	TMLE										Nearly unbiased with SL. Smaller bias than AIPTW when with GLM/GAM		Nominal coverage using bootstrap CI with GAM. Type I error < 5% with GAM/SL
	DSM		Slightly larger MSE than AIPTW/ TMLE			Slightly larger MSE than AIPTW/ TMLE	Nominal coverage and type I error ~5% with GAM	Nearly unbiased with GAM/SL. Smaller bias than AIPTW and TMLE when with GLM			Nearly unbiased with GLM/GAM/SL		Overcoverage and type I error < 5% with GLM/GAM/SL
	PENCOMP	Small bias with SL. Huge bias with GLM/GAM	Small MSE with SL. Huge MSE with GLM/GAM	Overcoverage with SL, undercoverage with GLM/GAM	Huge bias with GLM/GAM /SL	Small MSE with SL. Huge MSE with GLM/GAM	Undercoverage with GLM/GAM/SL	Small bias with SL. Huge bias with GLM/GAM	Small MSE with SL. Large MSE with GLM/GAM	Overcoverage with SL, undercoverage with GLM/GAM	Moderate bias with SL. Huge bias with GLM/GAM	Huge MSE with GLM/GAM /SL	Overcoverage with SL, undercoverage with GLM/GAM

Figure 6. Summary of the performance of various causal estimators under the four simulation scenarios. Colors indicate the performance of the estimator with green, yellow, and red meaning good, average, and poor performance, respectively.

Impact Questionnaire (FIQ), which is a continuous variable ranging from 0 to 80. There are 69 covariates in total, 24 of which are continuous variables and the other 45 are binary variables. Earlier studies showed there is no difference in FIQ among the treatment groups (Robinson et al. 2012; Yang et al. 2016). Here, we apply the compared causal estimators to estimate the causal effect of treatments for fibromyalgia on the FIQ score. The analysis steps are similar to those in Section 3.3. Figure 7 shows the performance of different estimators using different PS and outcome modeling. The estimated treatment effect is about 0.03 with confidence intervals of all doubly robust estimators including zero, which indicates there is no evidence that there are treatment effects between the OPI cohort and the TRA cohort on the FIQ score. With SuperLearner, the estimators achieve the smallest standard error.

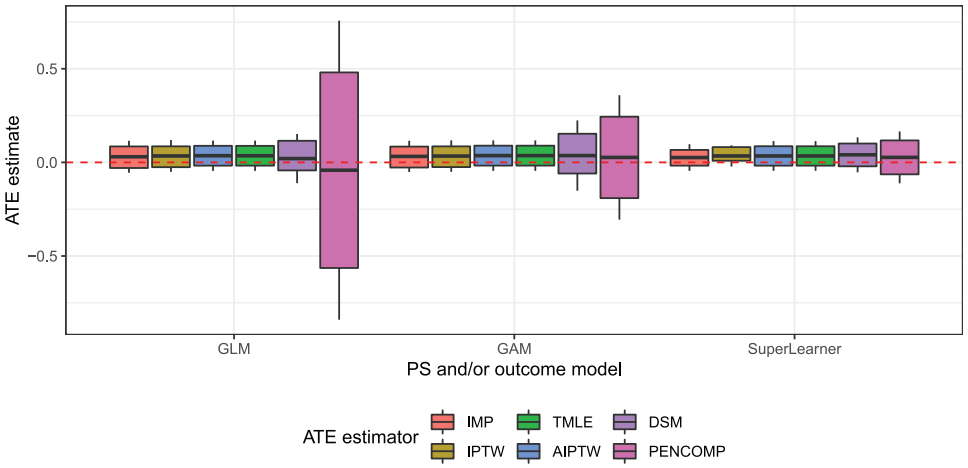


Figure 7. Performance of various causal estimators on the real-data application. Different colors imply different causal estimators, x-axis differentiate the PS and/or outcome models. The red dotted line indicates a zero ATE.

4.1. Simulated REFL study

Motivated by the real-world REFL study, we are interested in evaluating the effects of a higher dimension of covariates and the effects of binary variables on estimating treatment effects. The simulated REFL study is hence designed to mimic the data distribution in the real REFL study. We generate covariates X using Iman-Conover transformation (Iman and Conover 1982) to simulate correlated covariates from the real REFL data. The simulated REFL data is of sample size 2000 (1000 for the OPI cohort and 1000 for the TRA cohort).

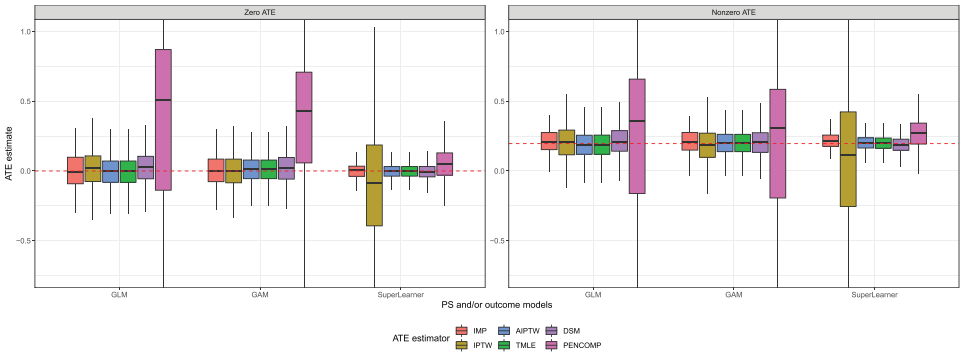
To simulate the treatment indicator A, we first generate the treatment model by fitting an XGBoost (Chen and Guestrin 2016) with cross-validation tuning using the real REFL data. This is referred to as the "true" treatment model. Specifically, the treatment indicator is used as the outcome and all patient characteristics are included as covariates. We then generate the simulated treatment assignment A for the simulated REFL data following Bernoulli($p^{\wedge}$) with $p^{\wedge}$ estimated from the "true" treatment model. Similarly, the simulated outcomes Y are obtained from the outcome model which is fit by an XGBoost with crossvalidation tuning. This is referred to as the "true" outcome model.

Two simulation scenarios are designed with different "true" outcome models. The first one is the zero treatment effect (Zero ATE) scenario, where the "true" outcome model is based on fitting an XGBoost to outcome data with no treatment indicator, then using the predicted value from that model, denoted as $Y^{\wedge}$, to generate simulated outcomes by adding a Gaussian noise with variation obtained from the cross-validation process. The resulting simulated outcomes are considered "observed" outcomes, and $Y^{\wedge}$'s are considered "truth". The second one is the nonzero treatment effect (Nonzero ATE) scenario, where the "true" outcome model is constructed by fitting an XGBoost to outcome data with A simulated from the "true" treatment model as a covariate and other covariates included in the candidate set. The corresponding predicted value $Y^{\wedge}$ from that fitted model is used to simulate data. There are 1000 simulated datasets generated, each one conducts a bootstrap 500 times.

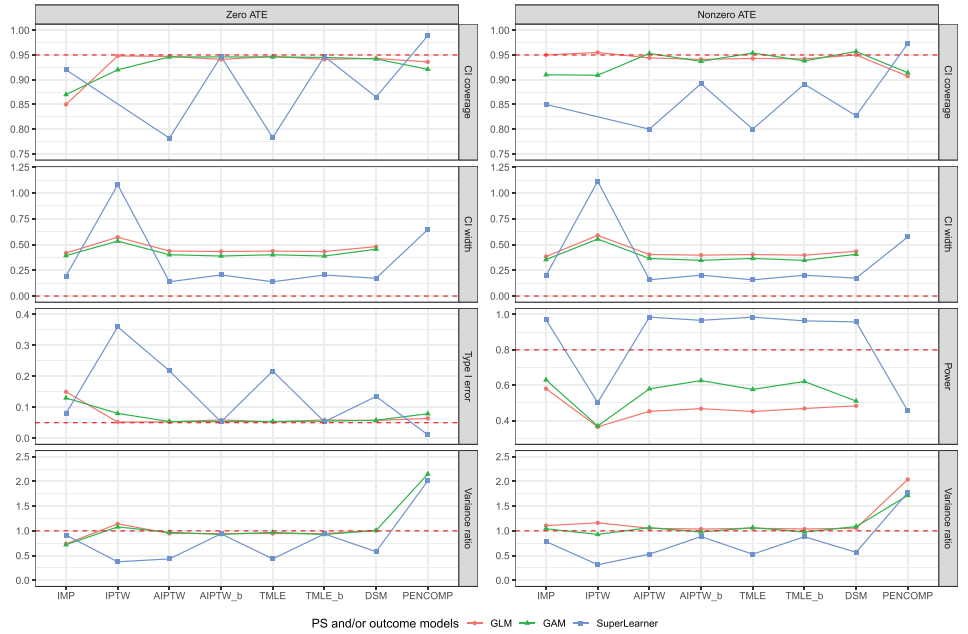
Figure 8 illustrates the performance of various causal estimators on the simulated REFL study considering Zero-ATE design and Nonzero-ATE design, respectively. Similar to findings in the simulation study in Section 3, doubly robust estimators, in general, outperform singly robust ATE estimators. Using SuperLearner for treatment and outcome modeling, doubly robust estimators achieve the smallest bias and MSE compared to using GLM, or GAM. Overall, TMLE and AIPTW share similar performance. With GAM or SuperLearner, TMLE and AIPTW may be able to achieve a nominal confidence interval by using bootstrap confidence intervals. DSM has a relatively larger bias compared to TMLESL, TMLE, and AIPTW. DSM may achieve a nominal confidence interval with GAM. PENCOMP suffers from severe bias and MSE with GLM or GAM. PENCOMP has improved performance with SuperLearner in terms of bias and MSE. However, PENCOMP suffers from severe overcoverage and large type I error in nearly all settings.

5. Practical recommendations and discussion

We have reviewed multiple doubly robust estimators and conducted simulations across a broad range of data scenarios. We vary causal inference test settings by adjusting a variety of knobs in the simulations, which include nonlinearity of treatment and outcome surfaces, degree of overlap between treatment distributions as well as treatment effect heterogeneity. We make use of a powerful machine learning technique SuperLearner to help improve ATE estimation. Also, various doubly robust estimators are applied to a real-life application of fibromyalgia as an example. In particular, we find that incorporating machine learning with doubly robust estimators such as the TMLE gives the best overall performance. Although in general TMLE and AIPTW are both efficient and have the minimum asymptotic variance under the large-sample theory, under finite sample sizes TMLE tends to be more robust to data sparsity and near violations of positivity assumption because of its range-preserving procedure for the predicted outcome estimates. Similar findings have been shown in previous studies such as van der Laan and Rose (2011), Porter et al. (2011), Luque-Fernandez et al. (2018), Bahamyirou et al. (2019). DSM is robust to model misspecification as a matching estimator, but tends to have a larger MSE compared to TMLE and AIPTW.



(a)



(b)

Figure 8. Performance of various causal estimators on the simulated REFL study considering both a homogeneous treatment effect design (left, zero ATE) and a heterogeneous treatment effect design (right, nonzero ATE), respectively. (a) Box plots of the estimators. The red dotted lines indicate the ground truth ATE. (b) Performance of the estimators in terms of CI coverage and width, type I error, and variance ratio. The red dotted lines indicate the ideal value or threshold of corresponding metrics. Methods with a "_b" indicate a bootstrap CI coverage.

The regression-based PENCOMP shows the least ideal performance among all doubly robust estimators in the case of model misspecification and challenging scenarios, even when pairing with SuperLearner. Further research is needed to demystify the performance of PENCOMP found in our simulation studies.

Our paper helps to provide guidelines for practical use of doubly robust estimators. Based on our extensive and realistic simulations, we recommend to estimate the ATE in the following steps:

- Perform variable selection to select outcome predictors.
- Model the PS and the outcomes with SuperLearner, separated by the treatment group and the control group.

- Estimate the ATE by applying TMLE with the estimated propensity and outcome estimates.
- Use bootstrap for variance estimation of the ATE.

Throughout the paper, we have found that machine learning methods, such as SuperLearner, improve performance compared to traditional approaches like GLM and GAM, particularly when there is uncertainty about the distributions of the propensity and outcome models. A practical recommendation is to apply machine learning techniques to both propensity and outcome modeling in doubly robust estimators. In real-world applications, it is crucial to balance model complexity and sample size. For smaller datasets, more complex methods can be useful, but careful consideration is needed to avoid overfitting. We recommend using simulations based on real-world data to identify the most appropriate method for varying sample sizes and to establish best practices for method selection.

This work has multiple limitations that should be noted. First, throughout the paper we only consider Lasso for variable selection to illustrate the importance of the variable selection procedure. There might be better ways to remove treatment predictors such as using machine learning algorithms like random forest and neural networks. Soft variable selection strategies may also be used where the variable selection is conducted without requiring any modeling on the outcome, and thus provides robustness against misspecification (Tang et al. 2021). Second, our work focuses only on doubly robust methods, which is in part due to their advantages in robustness over traditional methods. However, these methods still require the correct specification of at least one of the models. Recent research has proposed the use of model averaging across many methods to improve the robustness of comparative analyses (Zagar et al. 2022). Future work should compare the operating characteristics of doubly robust approaches to model averaging, or perhaps simply incorporating multiple doubly robust methods within the model averaging framework. In addition, an extension of evaluating the use of doubly robust estimators on survival data could be explored in the future. Another important direction for future research is the use of propensity score matching to build external control arms from real-world data, particularly in single-arm designs where the control arm may have a larger sample size than the treatment arm (e.g., multiple controls per treatment subject). Future studies could explore how different ratios of treatment to external control subjects affect estimator performance, particularly regarding bias and variance. Additionally, examining the robustness of estimators when external controls come from heterogeneous data sources would provide valuable insights for improving estimation in clinical trials, especially in rare diseases.

In summary, this work has provided best practice guidance on the use of doubly robust methods for comparative analysis based on real-world data. The use of machine learning for variable selection and model development, along with estimation of treatment effects using TMLE, is found to help improve operating characteristics of doubly robust methods.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

Abadie, A., and G. W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74 (1):235–267. doi: 10.1111/j.1468-0262.2006.00655.x.

Antonelli, J., M. Cefalu, N. Palmer, and D. Agniel. 2018. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics* 74 (4):1171–1179. doi: 10.1111/biom.12887.

Bahamyirou, A., L. Blais, A. Forget, and M. E. Schnitzer. 2019. Understanding and diagnosing the potential for bias when using machine learning methods with doubly robust causal estimators. *Statistical Methods in Medical Research* 28 (6):1637–1650. doi: 10.1177/0962280218772065.

Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61 (4):962–973. doi: 10.1111/j.1541-0420.2005.00377.x.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. 2006. Variable selection for propensity score models. *American Journal of Epidemiology* 163 (12):1149–1156. doi: 10.1093/aje/kwj149.

Cao, W., A. A. Tsiatis, and M. Davidian. 2009. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 96 (3):723–734. doi: 10.1093/biomet/asp033.

Chen, T., and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, Association for Computing Machinery, 785–794, (NY), NY, USA. doi: 10.1145/2939672.2939785.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. Double/Debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21 (1):C1–C68. doi:10.1111/ectj.12097.

Chipman, H. A., E. I. George, and R. E. McCulloch. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4 (1):266–298. doi: 10.1214/09-AOAS285.

Díaz, I., and M. J. van der Laan. 2011. Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics* 7 (1):1–20. doi:10.2202/1557-4679.1356.

Glynn, A. N., and K. M. Quinn. 2010. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis* 18 (1):36–56. doi: 10.1093/pan/mpp036.

Gruber, S., and M. van der Laan. 2012. Tmle: An r package for targeted maximum likelihood estimation. *Journal of Statistical Software* 51 (13):1–35. doi: 10.18637/jss.v051.i13.

Hansen, B. B. 2008. The prognostic analogue of the propensity score. *Biometrika* 95 (2):481–488. doi: 10.1093/biomet/asn004.

Hariton, E., and J. J. Locascio. 2018. Randomised controlled trials – the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics and Gynaecology* 125 (13):1716. doi: 10.1111/1471-0528.15199.

Iman, R. L., and W.-J. Conover. 1982. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation* 11 (3):311–334. doi: 10.1080/03610918208812265.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86 (1):4–29. doi: 10.1162/003465304323023651.

Kang, J. D. Y., and J. L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22 (4):523–539. doi: 10.1214/07-STS227.

Leacy, F. P., and E. A. Stuart. 2014. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: A simulation study. *Statistics in Medicine* 33 (20):3488–3508. doi: 10.1002/sim.6030 .

Little, R., and H. An. 2004. Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica* 14 (3):949–968. http://www.jstor.org/stable/24307424.

Little, R. J., and D. B. Rubin. 2019. *Statistical analysis with missing data*. (NJ): John Wiley & Sons.

Lunceford, J. K., and M. Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23 (19):2937–2960. doi: 10.1002/sim.1903.

Luque-Fernandez, M. A., M. Schomaker, B. Rachet, and M. E. Schnitzer. 2018. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine* 37 (16):2530–2546. doi: 10.1002/sim.7628.

Myers, J. A., J. A. Rassen, J. J. Gagne, K. F. Huybrechts, S. Schneeweiss, K. J. Rothman, M. M. Joffe, and R. J. Glynn. 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 174 (11):1213–1222. doi: 10.1093/aje/kwr364.

Naimi, A. I., A. E. Mishler, and E. H. Kennedy. 2023. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *American Journal of Epidemiology* 192 (9):1536–1544. doi: 10.1093/aje/kwab201.

Neyman, J. 1923. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. Excerpts reprinted (1990) in English. *Statistical Science* 5:463–472.

Otsu, T., and Y. Rai. 2017. Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association* 112 (520):1720–1732. doi: 10.1080/01621459.2016.1231613.

Pearl, J. 2011. Invited commentary: Understanding bias amplification. *American Journal of Epidemiology* 174 (11):1223–1227. doi: 10.1093/aje/kwr352.

Peters, J., D. Janzing, and B. Schölkopf. 2017. *Elements of causal inference: Foundations and learning algorithms*. (MA): The MIT Press.

Pirracchio, R., M. L. Petersen, and M. Van Der Laan. 2015. Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology* 181 (2):108–119. doi: 10.1093/aje/kwu253.

Porter, K. E., S. Gruber, M. J. Van Der Laan, and J. S. Sekhon. 2011. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics* 7 (1):1–34. doi: 10.2202/1557-4679.1308.

Prosperi, M., Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2 (7):369–375. doi: 10.1038/s42256-020-0197-y.

Robins, J. M., M. A. Hernan, and B. Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11 (5):550–560. doi: 10.1097/00001648-200009000-00011.

Robinson, R. L., K. Kroenke, P. Mease, D. A. Williams, Y. Chen, D. D'Souza, M. Wohlreich, and B. McCarberg. 2012. Burden of illness and treatment patterns for patients with fibromyalgia. *Pain Medicine* 13 (10):1366–1376. doi: 10.1111/j.1526-4637.2012.01475.x.

Rosenbaum, P. R. 2010. *Design of observational studies*, vol. 10, 978–1. (NY): Springer.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1):41–55. doi: 10.1093/biomet/70.1.41.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5):688–701. doi: 10.1037/h0037350.

Stolberg, H. O., G. Norman, and I. Trop. 2004. Randomized controlled trials. *American Journal of Roentgenology* 183 (6):1539–1544. doi: 10.2214/ajr.183.6.01831539.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25 (1):1–21. doi: 10.1214/09-STS313.

Tan, X. 2023. Causal inference under data restrictions. *arXiv preprint arXiv:2301.08788*.

Tan, X., J. Abberbock, P. Rastogi, and G. Tang. 2022. Identifying principal stratum causal effects conditional on a post-treatment intermediate response. Proceedings of the First Conference on Causal Learning and Reasoning, in Proceedings of Machine Learning Research 177: 734–753. https://proceedings.mlr.press/v177/tan22a.html.

Tan, X., C.-C. H. Chang, L. Zhou, and L. Tang. 2022. A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources. In Proceedings of the 39th International Conference on Machine Learning', Vol. 162 of Proceedings of Machine Learning Research, eds. K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, 21013–21036, PMLR. https://proceedings.mlr.press/v162/tan22a.html.

Tan, X., Z. Qi, C. Seymour, and L. Tang. 2022. RISE: Robust individualized decision learning with sensitive variables. Advances in Neural Information Processing Systems, 35, 19484-19498.

Tang, D., D. L. A. I. L. Kong, W. Pan, and L. Wang. 2021. Variable selection for doubly robust causal inference. *arXiv preprint arXiv:2007.14190*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard. 2007. Super learner. *Statistical Applications in Genetics and Molecular Biology* 6 (1):1–23. doi: 10.2202/1544-6115.1309.

van der Laan, M. J., and D. Rubin. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2 (1):1043–1083. doi: 10.2202/1557-4679.1043.

van der Laan, M., and S. Rose. 2011. *Targeted learning: Causal inference for observational and experimental data*. Springer Series in Statistics, Springer New York. https://books.google.com/books?id=RGnSX5aCAgQC.

Waernbaum, I. 2012. Model misspecification and robustness in causal inference: Comparing matching with doubly robust estimation. *Statistics in Medicine* 31 (15):1572–1581. doi: 10.1002/sim.4496.

Yang, S., and P. Ding. 2018. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* 105 (2):487–493. doi: 10.1093/biomet/asy008.

Yang, S., G. W. Imbens, Z. Cui, D. E. Faries, and Z. Kadziola. 2016. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* 72 (4):1055–1065. doi: 10.1111/biom.12505.

Yang, S., J. K. Kim, and R. Song. 2020. Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (2):445–465. doi: 10.1111/rssb.12354.

Yang, S., and Y. Zhang. 2022. Multiply robust matching estimators of average and quantile treatment effects. *Scandinavian Journal of Statistics* 50 (1):235–265. doi: 10.1111/sjos.12585.

Zagar, A., Z. Kadziola, I. Lipkovich, D. Madigan, and D. Faries. 2022. Evaluating bias control strategies in observational studies using frequentist model averaging. *Journal of Biopharmaceutical Statistics* 32 (2):247–276. doi: 10.1080/10543406.2021.1998095.

Zhang, G., and R. Little. 2009. Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics* 65 (3):911–918. doi: 10.1111/j.1541-0420.2008.01155.x.

Zhang, Y., S. Yang, W. Ye, D. E. Faries, I. Lipkovich, and Z. Kadziola. 2021. Practical recommendations on double score matching for estimating causal effects. *Statistics in Medicine* 15:1–25.

Zhao, H., and S. Yang. 2021. Outcome-adjusted balance measure for generalized propensity score model selection. *arXiv preprint arXiv:2107.12487. Journal of Statistical Planning and Inference* 221:188–200. doi: 10.1016/j.jspi.2022.04.004 .

Zhong, Y., E. H. Kennedy, L. M. Bodnar, and A. I. Naimi. 2021. AIPW an R package for augmented inverse probability–weighted estimation of average causal effects. *American Journal of Epidemiology*. URL: 190 (12):2690–2699. doi: 10.1093/aje/kwab207.

Zhou, J., Z. Zhang, Z. Li, and J. Zhang. 2015. Coarsened propensity scores and hybrid estimators for missing data and causal inference. *International Statistical Review* 83 (3):449–471. doi: 10.1111/insr.12082.

Zhou, T., M. R. Elliott, and R. J. Little. 2019. Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association* 114 (525):1–19. doi: 10.1080/01621459.2018.1518234.