Communications for Statistical Applications and Methods
2022, Vol. 29, No. 2, 177–191

https://doi.org/10.29220/CSAM.2022.29.2.177
Print ISSN 2287-7843 / Online ISSN 2383-4757

Exploring modern machine learning methods to improve causal-effect estimation

Yeji Kim[a], Taehwa Choi[a], Sangbum Choi[1,a]

[a]Department of Statistics, Korea University,Korea

Abstract

This paper addresses the use of machine learning methods for causal estimation of treatment effects from observational data. Even though conducting randomized experimental trials is a gold standard to reveal potential causal relationships, observational study is another rich source for investigation of exposure effects, for example, in the research of comparative effectiveness and safety of treatments, where the causal effect can be identified if covariates contain all confounding variables. In this context, statistical regression models for the expected outcome and the probability of treatment are often imposed, which can be combined in a clever way to yield more efficient and robust causal estimators. Recently, targeted maximum likelihood estimation and causal random forest is proposed and extensively studied for the use of data-adaptive regression in estimation of causal inference parameters. Machine learning methods are a natural choice in these settings to improve the quality of the final estimate of the treatment effect. We explore how we can adapt the design and training of several machine learning algorithms for causal inference and study their finite-sample performance through simulation experiments under various scenarios. Application to the percutaneous coronary intervention (PCI) data shows that these adaptations can improve simple linear regression-based methods.

Keywords: average causal effect, doubly-robust estimation, inverse probability weighting, propensity score, random forest, targeted learning

1. Introduction

In many clinical trials or social science studies, comparison of treatment effect between two groups is often regarded as an important part of the research. This effect can be measured by using difference or ratio between each treatment group, and average causal effect (ACE) is often of primary interest since individual treatment effect is difficult to obtain especially when the study is not randomized. Under conventional randomized experiments, the treatment effect can be easily evaluated, since all factors except the treatment group (confounding factor) can be controlled by experimenters. In an observational study, however, many confounding factors are present and they may distort the causal effect severely, because the distribution of covariates among subjects between treatments can be different (Rubin, 1978).

In such cases, the inverse probability weighting (IPW) method can be used to balance the propensity score distributions between two treatment groups by dividing the observed outcome with the estimated propensity scores (Austin, 2011). The resulting estimator may approximate the effect as if it were obtained from a randomized experiment, but it requires a well-posited propensity score (PS)

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea.
E-mail: choisang@korea.ac.kr

Published 31 March 2022 / journal homepage: http://csam.or.kr

model for stable analysis (Rosenbaum and Rubin, 1983). However, problems due to model misspecification often arise in real applications, since the true propensity model is hard to know exactly in an observational study. Alternatively, the augmented inverse probability weighting estimator (AIPWE), a hybrid approach of regression and IPW method, has been extensively studied as a generalization of the IPW method (Robins *et al.*, 1994; Kang and Schafer, 2007; Funk *et al.*, 2011). AIPWE is well-known to be doubly-robust (DR), because it still produces a consistent result if either the regression or IPW model is correctly specified (Tsiatis, 2007).

However, AIPWE could be worse than a single IPW method when both models are incorrectly posited (Kang and Schafer, 2007). There have been many studies to overcome the limitation of this DR estimator. In this case, machine learning techniques are a natural choice to accommodate potential model misspecification for causal regression modeling (Lee *et al.*, 2010; McCaffrey *et al.*, 2013). Gruber and van der Laan (2010) proposed a targeted maximum likelihood estimation (TMLE) that is another semiparametric DR method, equipped with nonparametric super-learner (SL) algorithms. Using only SL models can infer false confidence intervals if severe bias occurs and one of the two models is incorrectly specified. Therefore, TMLE takes a semiparametric approach to the SL model to compensate for this limitations to have an overall unbiased estimator at all levels (Schuler and Rose, 2017). Causal forest (CF) (Athey *et al.*, 2019) is a newly developed estimation method, based on ensemble models for causal inference. This method is developed from decision trees and random forest to estimate the treatment effect. It generates a large number of trees for each treatment group, then provides appropriate weights and variations for each group and finally generates an ACE estimator.

Despite various proposals for improving causal estimators, researchers may suffer from inadequate performance comparisons between these methods. In this article, we present an empirical study to compare several recently developed causal inference models for estimating ACE. We briefly overview recent developments in causal analysis, including regression-based estimator, IPWE, AIPWE, TMLE and CF and explore which methods perform better under various situations via simulation experiments. Our simulation study covers a wide range of scenarios that may change ACE and investigates existing and newly adapted machine learning models. We replace the regression with generalized smoothing model (GSM) and IPW models with several machine learning models with correct, misspecified and omitted models, and show how to have lower bias and average error rate in a wide range of scenarios.

The rest of the article is organized as follows. Section 2 presents basic notation and underlying assumptions and summarizes various recent causal effect estimating methods. Section 3 provides extensive simulation results that compare performance improvements with different combinations of the resulting regression and propensity score models. In Section 4, several causal models are applied to the percutaneous coronary intervention (PCI) data for illustration. Section 5 provides a short explanation and discussion of our results.

2. Methods

2.1. Preliminaries: Data and assumptions

In causal inference, we are often interested in clarifying the causal relationship in observational data. However, to allow for a precise presentation of the causal treatment estimators, we need additional notation and assumptions for the observed data (Hernán and Robins, 2020). For $i = 1, 2, \ldots, n$ individuals, let Y_i and A_i be the observed outcome and the treatment or exposure group variable, respectively, and X_i be the p-vector of observed pretreatment covariates. Then, the observed data are represented by $\{O_i = (X_i, A_i, Y_i),\ i = 1, \ldots, n\}$ and the data distribution follows an unknown distribution P. For

simplicity, we assume that A_i takes a binary value, $A_i \in \{0, 1\}$, but the method can be easily extended to the case with multiple treatments (McCaffrey *et al.*, 2013). Under the counterfactual framework (Rubin, 1978), each individual will have two potential outcomes, say $Y(a)$, which represents the outcome variable that corresponds to the exposure factor $A = a$. Suppose that we are mainly interested in inferring the average causal effect (ACE),

$$\tau \equiv E[Y(1) - Y(0)], \tag{2.1}$$

which measures average treatment difference between $Y(1)$ and $Y(0)$. However, the causal effect for an individual generally cannot be estimated because of the fundamental problem of causal inference: each individual will receive either $A = 1$ or 0, but not both at the same time.

To quantify the causal effect in observational studies, we shall make the following three common assumptions in causal analysis (Hernán and Robins, 2020),

(i) *consistency:* $Y = AY(1) + (1 - A)Y(0)$,

(ii) exchangeability: $Y(a) \perp\!\!\!\perp A \mid X$,

(iii) positivity: $0 < P(A = 1|X = x) < 1$ *for all* $x \in \mathcal{X}$.

The consistency assumption states that there are no multiple versions of treatment outputs, which means that the mechanism used to assign the treatments does not matter and assigning the treatments in a different way does not constitute a different treatment. The exchangeability assumption says the set of observed pretreatment covariates is sufficiently rich, such that there exist no unmeasured or unknown confounders, not included in X. Since there is no way to test this assumption with observed data, analysts should try to include a broad set of covariates to reduce the potential risk that a confounding variable is omitted inadvertently. Finally, the positivity assumption states the observed data for both treatments should be available for any given X, such that group comparison will make sense.

The commonly used randomized experimental designs allow researchers to estimate the average causal effect estimates closest to the theory, in which we may consider

$$\hat{\tau} = \frac{1}{\mid \{i : A_i = 1\} \mid} \sum_{\{i:A_i=1\}} Y_i - \frac{1}{\mid \{i : A_i = 0\} \mid} \sum_{\{i:A_i=0\}} Y_i.$$

Unlike randomized experimental data, however, the treatment group is unevenly distributed or associated with underlying confounders in observational data, making it difficult to use $\hat{\tau}$ for predicting the causal effect. There are several ways to solve this problem, such as matching, stratification of subclassification, adjustment by instrumental variables, and treatment weighting (Stuart, 2010; Robins *et al.*, 2007). In the following, we review a few recent developments for clarifying causal relationships between treatment and outcome in observational studies and provide some guidance on how we can implement machine learning techniques to improve the quality of causal estimates.

2.2. Classical approaches: Regression and inverse probability weighting (IPW) estimators

The estimation of treatment effects may proceed in two stages by modeling the conditional outcome $Q_a(x) = E[Y|A = a, X = x]$ and the propensity score $g(x) = P(A = 1|X = x)$. In principle, it is straightforward to use statistical regression models for $Q_a(x)$ and $g(x)$. First, we consider estimating

the average causal effect through regression models. Since the average effect under $A = a$ can be formulated by $\mu_a = E[Y|A = a] = \int_{\mathcal{X}} Q_a(x)P(dx)$, a regression-based causal effect estimate can be obtained by

$$\hat{\tau}^{\text{Reg}} = \hat{\mu}_1^{\text{Reg}} - \hat{\mu}_0^{\text{Reg}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{Q}_1(x_i) - \widehat{Q}_0(x_i)\right\}, \tag{2.2}$$

where $\widehat{Q}_a(x)$ is a regression estimate of the conditional mean outcome $Q_a(x)$. We can use standard parametric regression models or more flexible semiparametric or nonparametric methods to predict the outcome Y_i from the treatment and covariates.

In our simulations, we primarily employ a generalized smoothing model (GSM) by Helwig (2020), which estimates parameters by estimating the unknown smooth function under a multiple and generalized nonparametric regression model from the sample of data. The GSM algorithm is available in the R package, *npreg*, which offers several smoother types and will select the smoothing parameter basically based on generalized cross validation (GCV) according to the nominated predictor's class without specifying the model formula.

On the other hand, the inverse probability weight estimator (IPWE) can be described as a difference of weighted averages that assigns inverse probability weights to each treatment group. We can estimate the propensity score $g(x)$ by $\hat{g}(x)$ and obtain the treatment effect as

$$\hat{\tau}^{\text{IPWE}} = \hat{\mu}_1^{\text{IPWE}} - \hat{\mu}_0^{\text{IPWE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i = 1)}{\hat{g}(x_i)} - \frac{I(A_i = 0)}{1 - \hat{g}(x_i)}\right\} Y_i. \tag{2.3}$$

Conceptually, IPWE attempts to fully adjust for measured confounders by balancing the confounders across levels of treatment with treatment weight. When a function for the propensity score is unknown, logistic and multinomial regression models are commonly used for binary and multiple treatments, respectively.

The propensity scores, when properly estimated, ensure that covariates are assigned proportionally across two treatment groups, and may control very distant outliers. In this case, machine learning methods, such as boosted ensemble models, have been shown to yield causal effect estimates with many desirable properties (Lee *et al.*, 2010; McCaffrey *et al.*, 2013). We shall also explore generalized boosted model (GBM) for estimation of the necessary propensity score weights.

2.3. Augmented inverse probability weighting estimator (AIPWE)

An alternative and improved estimator is the augmented inverse probability weighted estimator (AIPWE) that combines the properties of both regression-based estimator and IPWE (Robins *et al.*, 1994; Kang and Schafer, 2007; Funk *et al.*, 2011). AIPWE requires two models: the outcome model and the propensity model. The outcome model typically uses a regression model and must specify an outcome model formula that states the relationship between covariates and outcome variables within each treatment group. Likewise, the propensity model specifies a model that relates covariate effects with treatment allocation. Specifically, the AIPWE takes the form

$$\hat{\tau}^{\text{AIPWE}} = \hat{\mu}_1^{\text{AIPWE}} - \hat{\mu}_0^{\text{AIPWE}}, \tag{2.4}$$

where

$$\hat{\mu}_1^{\text{AIPWE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i=1)Y_i}{\hat{g}(x_i)} - \frac{I(A_i=1)-\hat{g}(x_i)}{\hat{g}(x_i)}\widehat{Q}_1(x_i)\right\},$$
$$\hat{\mu}_0^{\text{AIPWE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i=0)Y_i}{1-\hat{g}(x_i)} + \frac{I(A_i=0)-\hat{g}(x_i)}{1-\hat{g}(x_i)}\widehat{Q}_0(x_i)\right\}.$$

Notice that the first term in $\hat{\mu}_a^{\text{AIPWE}}$ $(a = 0, 1)$ is equivalent to $\hat{\mu}_a^{\text{IPWE}}$ in (2.3), while the second term augments the IPWE to reduce variability and improve estimate efficiency by appending the regression estimator $\widehat{Q}_a(x)$ in a structured way. The AIPWE is known to be "doubly robust" in that it only requires either the propensity or outcome model to be correctly specified but not necessarily both (Tsiatis, 2007). The "augmented" term could complement the bias with zero even when either the propensity model or the outcome model goes wrong. AIPWE can also be aided by machine learning methods; for example, we may use GBM for the propensity score model instead of standard logistic model and GSM for the outcome model instead of linear regression model. In the last decade, AIPWE has gained much popularity and has been extensively studied due to its efficiency, reliability and simple implementation. Recently, Choi *et al.* (2021) proposed a pseudo-value-based AIPWE when the response is subject to censoring and showed that the double-robustness property is still valid under model misspecification.

2.4. Targeted maximum likelihood estimator (TMLE)

TMLE (van der Laan and Rose, 2011; Schuler and Rose, 2017) is another well-established doubly-robust maximum likelihood method for causal effect estimation. Like AIPWE, the TMLE procedure should also specify both the outcome and propensity models. It begins by estimating the conditional mean models given the treatment and covariates, $Q_1(x)$ and $Q_0(x)$, and the propensity score model $g(x)$ with a prespecified method. Then one can create so-called "clever" covariate, which is very similar to the IPWE format and given by

$$H(x) = H_1(x) - H_0(x) = \frac{I(A=1)}{\hat{g}(x)} - \frac{I(A=0)}{1-\hat{g}(x)}.$$

Then, we generate "targeted" estimates of the set of potential outcomes, incorporating the clever covariate as an adjustment to reduce the bias. Specifically, by treating the initial estimate $\widehat{Y}_a = \widehat{Q}_a(x)$ as an offset, fit the following model

$$h\left(E[Y|A=a, X=x]\right) = h\left(\widehat{Y}_a\right) + \delta H(x),$$

to obtain the estimate $\hat{\delta}$ of the fluctuation parameter δ. $h(\cdot)$ denotes a link function in generalize linear model. Note that $H(x)$ is the term to compensate for the estimation of the outcome model and relies on the variation of the parameters of interest and $\hat{\delta}$ is the maximum likelihood estimator obtained by regressing $H(x)$ on Y with the offset $\widehat{Y}_a$. Finally, we can estimate the ACE by

$$\hat{\tau}^{\text{TMLE}} = \hat{\mu}_1^{\text{TMLE}} - \hat{\mu}_0^{\text{TMLE}}, \tag{2.5}$$

where $\hat{\mu}_a^{\text{TMLE}} = 1/n\sum_{i=1}^{n}\widehat{Q}_a^*(x_i)$, $(a = 0, 1)$. Here, $\widehat{Q}_a^*(x)$ can be obtained from $h(\widehat{Q}_a^*(x)) = h(\widehat{Y}_a) + \hat{\delta}\, H(x)$.

TMLE is a two-step semiparametric procedure that solves the efficient influence curve estimating equation and thereby yields an efficient or at least locally efficient solution to the parameter of interest. In addition, combined with the super-learner (SL) algorithm (van der Laan *et al.*, 2007), TMLE has demonstrated its utility as a very powerful tool for efficient substitution estimator. By using machine learning methods, TMLE can have many advantages, such that it can reflect the underlying features of data smoothly needless of assuming restrictive conditions on a functional relationship of the outcome and exposure mechanisms. For example, several ensemble methods are considered to achieve better performance and then one of them can be selected according to data based on cross-validation and the ensemble algorithm can be applied to the data simultaneously and synergistically. Super learner is the final weighted results aggregation that combines several ensemble learnings by their performance, and is known to work very well with minimal cross-validated mean squared error. It can be implemented with the *tmle* package in R and statistical inference can be made by using the estimator's influence curve or bootstrapping.

2.5. Nonparametric causal forest method

Causal forest (CF) is a causal inference learning method as an extension of Breiman (2001)'s random forest (Wager and Athey, 2018; Athey *et al.*, 2019). In random forest, data are repeatedly split in order to minimize prediction error of an outcome variable. Causal forest is built similarly, except that instead of minimizing prediction error, data are split in order to maximize the difference across splits in the relationship between outcome variable and treatment variable. This is intended to uncover how treatment effects vary across a sample. Additionally, it can exploit the large feature space characteristic of big data and abstract inherent heterogeneity in treatment effects.

Fundamental idea of CF is constructing partitions of covariate act as sample came from randomized trial, which can be implemented as follows. Suppose there exist L partitions of covariate space (denoted by L-leaves) given an existing split rule. Furthermore, the leaf $L(x)$ with test covariate value x is assumed to sufficiently small so that (Y_i, A_i) in $i \in L(x)$ behave as they were gathered from a randomized trial. One can approximate the average treatment effect for any $x \in L$ as

$$\hat{\mu}_a^{\text{Tree}}(x) = \frac{1}{\mid \{i : A_i = a, X_i \in L(x)\} \mid} \sum_{\{i:A_i=a, X_i \in L(x)\}} Y_i. \tag{2.6}$$

Then, the corresponding ACE estimate with causal tree can be obtained by $\hat{\tau}^{\text{Tree}}(x) = \hat{\mu}_1^{\text{Tree}}(x) - \hat{\mu}_0^{\text{Tree}}(x)$. Likewise other random forest, one can use a bagging algorithm, aggregating the results of multiple causal trees from bootstrapped samples. Denote $\hat{\tau}_b^{\text{Tree}}$ as an ACE estimate from the bth causal tree with bootstrapped samples for $b = 1, \ldots, B$. Then, CF aggregates $\hat{\tau}_b^{\text{Tree}}$ and yields the final causal estimator by averaging their scores, that is,

$$\hat{\tau}^{\text{CF}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b^{\text{Tree}}. \tag{2.7}$$

The CF algorithm is available in the R package, *grf*, where double-sample (validation data and training data) tree is the default base model; Athey *et al.* (2019) for further applications.

Unlike conventional tree algorithms, CF algorithm is not based on minimizing the mean squared error in a regression tree. Since the true τ is not observable, we cannot calculate the difference of $\hat{\tau} - \tau$. We instead calculate the conditional average treatment effect (CATE) for covariates in each final leaf L as (2.6) and (2.7). According to Athey and Imbens (2016), one can optimize CF algorithm

by maximizing the variance of $\hat{\tau}(x)$ rather than minimizing the mean squared error in a regression tree. Therefore, if the variance of $\hat{\tau}(x)$ is maximized during the splitting step, the value of the splitting criteria can be selected as an optimal value (Gulen *et al.*, 2020). Although the variance can increase due to smaller sample sizes, causal estimates can be free of overfitting and bias, as in random forest. Further, one can control the increased variance to some extent by more repeating the bagging procedure (2.7). According to Wager and Athey (2018), $\hat{\tau}(x)$ is a consistent, asymptotically normal estimate of CATE.

3. Simulation studies

In this section, we provide empirical simulation results to evaluate the finite-sample performance of several causal inference methods under a wide range of scenarios. For each outcome model (m) and propensity score model (π), we consider three scenarios: (i) correct-model specification, (ii) incorrect-model specification and (iii) omitted-model specification. This setting yields 9 possible combinations of outcome and propensity models. Each model also involves some nonlinear and interaction terms, which are included to explore the performance of machine learning algorithms for causal inference in more realistic settings.

Our simulation involves three covariates, $Z = (Z_1, Z_2, Z_3)'$, which follows a multivariate normal distribution with mean zero, unit variance and $\text{Cor}(Z_i, Z_j) = \pm 0.2$, $(i \neq j)$. The correct-model case involves the correct outcome variable

$$Y = 5 + Z_1^2 + Z_1 Z_2 + Z_2 + Z_3 + A\,(2 + Z_1 + Z_2 + Z_3) + \epsilon,$$

where $\epsilon \sim N(0, 1)$. The treatment variable $A \in \{0, 1\}$ is generated from a logistic regression model with $P(A = 1|Z) = \text{expit}(Z_1 + 0.7Z_1 I(Z_2 > 0) + Z_3)$ as P_1 and $P(A = 1|Z) = \text{expit}(Z_1 + 3.5Z_1 I(Z_2 > 0) + Z_3)$ as P_2, respectively, where $\text{expit}(x) = 1/(1 + e^{-x})$. Notice that P_2 has a greater magnitude of nonlinearity than P_1. Under the incorrect-model case, we fit the outcome variable Y with A and covariates $X = (X_1, X_2, X_3)$ in place of $Z = (Z_1, Z_2, Z_3)$ in both outcome model (m) and IPW model (π), where

$$X_1 = \exp(-2Z_3), \quad X_2 = Z_1 I(Z_2 > 0) - Z_3 I(Z_2 < 0), \quad X_3 = Z_3 + Z_1 Z_2.$$

In the omitted model case, we assume that Z_2 is wrongly omitted in the outcome model (m) and IPW model (π), resulting in fitting the outcome variable Y with treatment variable A and covariates (Z_1, Z_3) instead of $Z = (Z_1, Z_2, Z_3)$ in outcome model (m) and treatment variable A with covariates (Z_1, Z_3) in IPW model (π).

In this setting, we seek to find which of the proposed estimators provide higher robustness to model disturbances under nine scenarios, where both or only one of outcome model and IPW model are accurately or incorrectly specified or omitted. For ease of presentation, we consider two outcome models, linear regression (Reg) and GSM, with different specifications in outcome model (m). We cover the IPW model (π) with a logistic model (GLM) and three machine learning methods, including tree (Tree), gradient boosting machine (GBM), and super learner (SL). The SL model consists of predictions from five base learners, (i) *ipredbagg*, (ii) *ksvm*, (iii) *polymars*, (iv) *rpartPrune*, and (v) *KernelKnn*, to provide a significant opportunity for improvement in predictive power. The base learners are fundamental factors of the SL ensemble model, which are combined to enhance the predictive performance synergetically. Each of these base learners are available in the following R packages;

(i) improved predictive models by indirect classification and bagging for classification, regression and survival problems as well as resampling based estimators of prediction error (*ipred*), (ii) kernel-based machine learning methods for classification, regression, clustering, novelty detection, quantile regression and dimensionality reduction (*kernlab*), (iii) polynomial spline routines for the polynomial

Table 1: Summary statistics of simulation results for $\tau = \mu_1 - \mu_0$ under the correct model (C), misspecified model (M), and omitted model (O) scenarios. Linear regression (Reg) and generalized smoothing model (GSM) are used for outcome model (m), while logistic regression (GLM), decision tree (Tree), generalized boosting model (GBM), and super-learner (SL) algorithms are used for propensity score model (π). Sample size is $N = 500$ and $P_1 = \text{expit}(Z_1 + 0.7Z_1I(Z_2 > 0) + Z_3)$

	Model			GLM		Tree		GBM		SL	
Method	m	π	Q	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
TMLE	C	C	GSM	0.002	0.026	0.005	0.023	0.005	0.023	0.005	0.020
			Reg	–0.008	0.271	–0.006	0.053	–0.004	0.062	–0.001	0.039
	C	M	GSM	0.001	0.061	0.002	0.024	0.004	0.020	0.003	0.020
			Reg	0.018	0.169	–0.002	0.051	0.001	0.042	0.000	0.041
	C	O	GSM	0.003	0.025	0.005	0.022	0.005	0.023	0.004	0.020
			Reg	–0.006	0.265	–0.006	0.053	–0.004	0.066	–0.003	0.040
	M	C	GSM	–0.095	0.142	0.011	0.092	–0.021	0.071	0.084	0.067
			Reg	–0.130	0.178	0.025	0.099	–0.022	0.078	0.113	0.070
	M	M	GSM	0.545	0.616	0.259	0.151	0.313	0.166	0.268	0.135
			Reg	0.689	0.847	0.295	0.177	0.348	0.189	0.314	0.160
	M	O	GSM	–0.094	0.173	0.008	0.095	–0.036	0.112	0.050	0.084
			Reg	–0.129	0.214	0.019	0.099	–0.043	0.125	0.068	0.084
	O	C	GSM	0.204	0.098	0.160	0.093	0.193	0.083	0.133	0.060
			Reg	0.209	0.337	0.122	0.104	0.169	0.105	0.097	0.064
	O	M	GSM	–0.347	0.519	–0.001	0.088	0.005	0.057	0.054	0.051
			Reg	–0.363	0.624	–0.037	0.114	–0.032	0.079	0.016	0.067
	O	O	GSM	0.204	0.121	0.186	0.103	0.196	0.109	0.188	0.097
			Reg	0.212	0.356	0.149	0.111	0.176	0.133	0.153	0.095
AIPWE	C	C	GSM	0.001	0.027	0.004	0.023	0.003	0.022	0.003	0.020
			Reg	–0.006	0.321	0.001	0.054	0.002	0.063	0.005	0.040
	C	M	GSM	0.005	0.147	0.001	0.024	0.002	0.020	0.002	0.020
			Reg	0.051	0.533	0.001	0.052	0.002	0.042	0.003	0.041
	C	O	GSM	0.001	0.027	0.002	0.068	0.004	0.023	0.003	0.020
			Reg	–0.004	0.314	0.002	0.053	0.002	0.068	0.004	0.040
	M	C	GSM	–0.236	0.281	–0.126	0.191	–0.133	0.167	–0.020	0.146
			Reg	–0.227	0.273	–0.065	0.135	–0.081	0.114	0.066	0.092
	M	M	GSM	0.513	1.134	0.131	0.186	0.188	0.185	0.149	0.171
			Reg	0.733	1.579	0.222	0.171	0.280	0.176	0.258	0.158
	M	O	GSM	–0.234	0.315	–0.131	0.198	–0.152	0.215	–0.058	0.171
			Reg	–0.225	0.312	–0.073	0.139	–0.106	0.167	0.015	0.110
	O	C	GSM	0.203	0.106	0.159	0.097	0.193	0.083	0.140	0.062
			Reg	0.203	0.387	0.126	0.106	0.169	0.105	0.106	0.066
	O	M	GSM	–0.414	1.493	–0.004	0.094	0.015	0.057	0.073	0.054
			Reg	–0.427	1.966	–0.044	0.119	–0.029	0.079	0.031	0.068
	O	O	GSM	0.204	0.126	0.185	0.107	0.195	0.109	0.188	0.096
			Reg	0.206	0.404	0.153	0.113	0.175	0.133	0.156	0.096
IPWE		C		0.209	0.717	0.544	0.395	0.249	0.243	0.313	0.182
		M		1.249	6.745	0.424	0.296	0.365	0.254	0.295	0.173
		O		0.213	0.713	0.590	0.445	0.238	0.264	0.395	0.249

spline fitting routines hazard regression, hazard estimation (*polspline*), (iv) recursive partitioning for classification, regression and survival trees (*rpart*), and (v) the extended simple k-nearest neighbors algorithm by incorporating numerous kernel functions and a variety of distance metrics (*KernelKnn*).

Tables 1 and 2 show the simulation results of IPWE, AIPWE and TMLE estimators in the correct, incorrect and omitted model specification of the outcome and IPW models, respectively, to estimate the average causal effect estimator (ACE). Simulation results in Tables 1 and 2 are based on the

Table 2: Summary statistics of simulation results for $\tau = \mu_1 - \mu_0$ under the correct model (C), misspecified model (M), and omitted model (O) scenarios. Linear regression (Reg) and generalized smoothing model (GSM) are used for outcome model (m), while logistic regression (GLM), decision tree (Tree), generalized boosting model (GBM), and super-learner (SL) algorithms are used for propensity score model (π). Sample size is $N = 500$ and $P_2 = \text{expit}(Z_1 + 3.5Z_1I(Z_2 > 0) + Z_3)$

	Model			GLM		Tree		GBM		SL	
Method	m	π	Q	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
TMLE	C	C	GSM	–0.007	0.044	–0.004	0.033	–0.004	0.031	–0.003	0.028
			Reg	–0.052	1.566	–0.022	0.077	–0.017	0.090	–0.006	0.055
	C	M	GSM	–0.005	0.042	–0.003	0.034	–0.003	0.027	–0.004	0.026
			Reg	0.005	0.126	–0.003	0.071	0.001	0.053	0.002	0.053
	C	O	GSM	–0.007	0.044	–0.005	0.029	–0.004	0.032	–0.005	0.027
			Reg	–0.051	1.536	–0.023	0.072	–0.017	0.092	–0.014	0.055
	M	C	GSM	–0.480	0.625	–0.036	0.139	–0.211	0.146	0.033	0.069
			Reg	–0.547	0.711	–0.028	0.133	–0.219	0.148	–0.146	0.126
	M	M	GSM	0.505	0.391	0.360	0.246	0.362	0.202	0.317	0.170
			Reg	0.637	0.626	0.389	0.284	0.390	0.219	0.309	0.171
	M	O	GSM	–0.474	0.640	–0.127	0.118	–0.241	0.190	–0.084	0.096
			Reg	–0.542	0.727	–0.126	0.117	–0.254	0.193	–0.186	0.127
	O	C	GSM	0.495	0.357	0.342	0.226	0.439	0.253	0.312	0.152
			Reg	0.619	2.353	0.241	0.178	0.390	0.256	0.082	0.108
	O	M	GSM	–0.299	0.320	–0.060	0.157	0.109	0.077	0.186	0.098
			Reg	–0.368	0.425	–0.140	0.197	0.024	0.087	–0.061	0.099
	O	O	GSM	0.497	0.375	0.436	0.266	0.446	0.278	0.443	0.266
			Reg	0.622	2.353	0.350	0.226	0.403	0.283	0.351	0.216
AIPWE	C	C	GSM	–0.008	0.059	–0.004	0.033	–0.005	0.030	–0.004	0.027
			Reg	0.000	3.096	–0.008	0.077	–0.006	0.091	–0.003	0.056
	C	M	GSM	–0.002	0.083	–0.004	0.034	–0.005	0.026	–0.005	0.026
			Reg	0.023	0.362	–0.006	0.074	–0.004	0.054	–0.003	0.054
	C	O	GSM	–0.007	0.059	–0.006	0.029	–0.005	0.031	–0.005	0.027
			Reg	0.001	2.989	-0.005	0.072	–0.007	0.094	–0.003	0.056
	M	C	GSM	–0.662	1.395	–0.114	0.177	–0.256	0.199	0.026	0.088
			Reg	–0.758	1.573	–0.107	0.162	–0.264	0.195	0.059	0.072
	M	M	GSM	0.494	0.555	0.310	0.233	0.312	0.190	0.301	0.166
			Reg	0.671	1.015	0.350	0.264	0.355	0.201	0.329	0.180
	M	O	GSM	–0.652	1.379	–0.217	0.174	–0.294	0.253	–0.112	0.123
			Reg	–0.750	1.556	–0.221	0.164	–0.309	0.252	–0.103	0.102
	O	C	GSM	0.503	0.428	0.344	0.229	0.442	0.261	0.336	0.168
			Reg	0.623	3.915	0.246	0.184	0.387	0.257	0.245	0.133
	O	M	GSM	–0.355	0.785	–0.074	0.169	0.132	0.087	0.228	0.115
			Reg	–0.460	1.182	–0.184	0.232	0.025	0.089	0.127	0.102
	O	O	GSM	0.505	0.444	0.436	0.267	0.448	0.283	0.444	0.266
			Reg	0.627	3.842	0.356	0.232	0.399	0.282	0.368	0.221
IPWE		C		0.670	5.469	0.821	0.852	0.418	0.405	0.281	0.198
		M		0.970	4.260	0.567	0.551	0.457	0.321	0.350	0.221
		O		0.672	5.256	0.938	1.000	0.417	0.403	0.576	0.436

previously described model specifications with P_1 and P_2, respectively. The data sample size is fixed at $N = 500$. The true parameter τ is approximated with pseudo-random estimates with sample size $N = 10^6$, and given by $\hat{\tau} = 1.996$ for Table 1 and $\hat{\tau} = 1.999$ for Table 2. Overall, the robust estimators, such as AIPWE and TMLE, perform well on finite sample sizes. It can be seen that AIPWE and TMLE have usually small biases when at least one model is correctly specified. Compared to naïve GLM, ensemble methods can significantly reduce both measures of bias and variance, contributing

Table 3: Simulation results of regression model with GLM (REG+), causal forest (CF), TMLE-SL and AIPWE-SL. Bias and MSE (in parentheses) are calculated, based on the sample size $N = 250$, $N = 500$ and $N = 1000$ with the prespecified propensity models, P_1 and P_2. Model specifications (m, π) are just applied to TMLE and AIPWE. CF are affected by only regression model specification (m). In each simulation, the best models in terms of MSE are marked as in bold

		Model		Method			
PS	*N*	*m*	π	REG+	CF	TMLE	AIPWE
P_1	250	*C*	*C*	0.009 (0.105)	0.251 (0.120)	0.013 (0.041)	**0.011 (0.040)**
		M	*C*	–0.020 (0.161)	0.252 (0.155)	**0.113 (0.124)**	–0.021 (2.530)
		O	*C*	0.167 (0.148)	0.226 (0.162)	**0.144 (0.106)**	0.151 (0.113)
	500	*C*	*C*	0.002 (0.053)	0.071 (0.031)	0.005 (0.020)	**0.003 (0.020)**
		M	*C*	–0.034 (0.079)	0.146 (0.069)	**0.084 (0.067)**	–0.020 (0.146)
		O	*C*	0.154 (0.084)	0.157 (0.083)	**0.133 (0.060)**	0.140 (0.062)
	1000	*C*	*C*	0.003 (0.027)	0.042 (0.013)	**0.006 (0.010)**	**0.006 (0.010)**
		M	*C*	**–0.051 (0.032)**	0.131 (0.040)	0.079 (0.038)	0.038 (0.093)
		O	*C*	0.156 (0.056)	0.164 (0.055)	0.139 (0.068)	**0.143 (0.042)**
P_2	250	*C*	*C*	0.005 (0.133)	0.268 (0.137)	0.009 (0.055)	**0.007 (0.054)**
		M	*C*	–0.222 (0.210)	0.240 (0.154)	**0.075 (0.138)**	0.019 (0.252)
		O	*C*	0.372 (0.306)	0.427 (0.309)	**0.319 (0.208)**	0.342 (0.226)
	500	*C*	*C*	–0.006 (0.073)	–0.046 (0.034)	**–0.003 (0.028)**	–0.004 (0.027)
		M	*C*	–0.247 (0.132)	**0.112 (0.066)**	0.033 (0.069)	0.026 (0.088)
		O	*C*	0.362 (0.222)	0.374 (0.206)	**0.312 (0.152)**	0.336 (0.168)
	1000	*C*	*C*	0.001 (0.036)	–0.067 (0.019)	0.006 (0.014)	**0.005 (0.014)**
		M	*C*	–0.260 (0.101)	**0.093 (0.035)**	0.015 (0.036)	0.032 (0.045)
		O	*C*	0.368 (0.181)	0.391 (0.182)	**0.321 (0.128)**	0.344 (0.143)

to MSE. When both models are misspecified or informative covariates are wrongly omitted in both models, biases cannot be completely removed, but still machine learning methods help effectively reduce bias and MSE. In Table 2, we found that AIPW and TMLE work stably well even when A is severely unbalanced. Note that AIPW and TMLE aided with SL algorithm show the best performance in nearly all cases.

Table 3 summarizes the results of CF, TMLE-SL and AIPWE-SL estimators. Overall, CF provides a great performance without requiring any complex model specifications. However, when both outcome and propensity models are correctly specified, CF seems to slightly underperform TMLE-SL and AIPWE-SL with respect to both bias and MSE. With small samples, CF has a relatively large MSE, but as the sample size gradually increases, the difference gradually decreases and in some case, CF shows a better MSE performance. However, we note that the simulation scenarios in Table 3 is not comprehensive and care should be taken for generalizing the previous findings.

Finally, Figure 1 displays eleven ACE estimators when both the outcome and propensity models are misspecified. See the note of Figure 1 for model description: for example, "model 4: TMLE-Reg-SL" denotes the composite TMLE method by implementing Reg for the outcome model and SL for the propensity model. Notice that CF (model 1) performs the best in this case, while two IPWEs (model 2 and 3) clearly underperform the other competitors due to model misspecification. Although AIPWE and TMLE work well without much difference, their performance depends on propensity model specifications rather than outcome models. For example, AIPWE-Reg-SL (model 5) outperforms AIPWE-Reg-GLM (model 7), while the difference between AIPWE-Reg-SL (model 5) and AIPWE-GSM-SL (model 8) is not much noticeable. On the other hand, by comparing IPWE-SL (model 2) and AIPWE-Reg-SL (model 5), we can see that doubly-robust estimation helps reduce both bias and variance even though model misspecification is present. We can conclude that the use of flex-

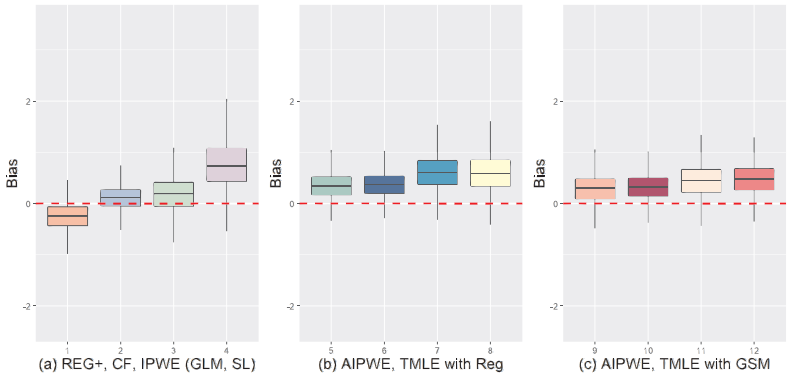

Figure 1: *Empirical comparison of several ACE estimators with misspecified π and misspesified m: (1) REG+, (2) CF, (3) IPWE-SL, (4) IPWE-GLM, (5) AIPWE-Reg-SL, (6) TMLE-Reg-SL, (7) TMLE-Reg-GLM, (8) AIPWE-Reg-GLM, (9) AIPWE-GSM-SL, (10) TMLE-GSM-SL, (11) AIPWE-GSM-GLM, and (12) TMLE-GSM-GLM.*

ible machine learning methods, in particular for propensity models, is beneficial in reducing potential biases and achieving more robust results to model misspecification.

4. Data example

As an illustration, we consider the data from a percutaneous coronary intervention (PCI) observational study with 996 patients at Ohio Heart Health, Christ Hospital, Cincinnati, in 1997 (Li and Shen, 2020). In the study, researchers continued to support Lindner Center for at least six months using two treatment methods, *abciximab* (expensive high molecular weight IIb/IIIa cascade blocker) for 70% of patients and general treatment alone with initial PCI for 29.9% of patients. The dataset has ten variables from 996 patients without omission. We consider *mortality* and *cardbill* as response variables; *mortality* represents life years, having 0 or 11.6 depending on whether died within 6 months or not and *cardbill* represents the total cost of treating a patient's heart within the first PCI 12 months at the Lindner Center. The *abcix* variable indicates whether the patient is treated *abciximab* or normal treatment alone. In addition, there are seven confounding variables; *stent*, whether to use a stent (anti-collapse device) or not; *height*, ranging from 108 cm to 196 cm; *sex*; *diabetic state*, whether to be diabetic or not; *acutemi*, whether the patient suffered from acute myocardial infarction within the last 7 days; *ejecfrac*, the ratio of left ventricle ejection from 0% to 90%; and *ves1proc*, the number of vessels involved in the initial PCI of patients from 0 to 5. Our main objective is to approximate the causal effects of the treatment variable *abcix* on *mortality* or *cardbill*. We apply several causal inference estimation methods to PCI data to determine whether the treatment is cost-effective after calibrating confounders and ensuring that there are more survivors of *abcix* treatment than people with other treatment. We estimate the bootstrap standard error by generating 10,000 bootstrap resamples. Although nearly 70% of patients are treated with *abcix* treatment, it is not clear that *abcix* treatment are reasonably inexpensive and actually makes people live more than others. Given only the numerical value of PCI data, the cardbill for *abcix* treatment is much more expensive than general treatment alone and survivors of each group are similar.

We begin by estimating propensity scores using logistic, Tree, GBM, and SL algorithms, including seven clinical predictors. Figure 2 displays the propensity score distributions for *abcix* treatment and normal treatment under four approaches, which shows the two distributions for each treatment group

Table 4: Summary results of PCI data analysis from regression (Reg), IPWE, AIPWE, TMLE and CF methods

Method	Model		Outcome	ACE	Bootstrap SE	Wald 95% CI
	m	π				
Reg	Reg		mortality	–0.050	0.002	(–0.054, –0.046)
			cardbill	981.579	96.361	(792.712, 1170.446)
	GSM		mortality	–0.051	0.002	(–0.055, –0.046)
			cardbill	1052.146	199.176	(661.761, 1442.531)
IPWE		GLM	mortality	–0.067	0.032	(–0.130, –0.003)
			cardbill	–27.852	1980.224	(–3909.092, 3853.388)
		Tree	mortality	–0.033	0.015	(–0.063, –0.004)
			cardbill	1537.173	1411.401	(–1229.174, 4303.520)
		GBM	mortality	–0.049	0.023	(–0.093, –0.005)
			cardbill	1690.826	1556.244	(–1359.413, 4741.065)
		SL	mortality	–0.032	0.014	(–0.059, –0.005)
			cardbill	1844.531	1359.656	(–820.394, 4509.456)
AIPWE	Reg	GLM	mortality	–0.065	0.027	(–0.118, –0.012)
			cardbill	251.458	1123.174	(–1949.963, 2452.879)
		Tree	mortality	–0.042	0.014	(–0.070, 0.013)
			cardbill	1262.021	829.256	(–363.320, 2887.363)
		GBM	mortality	–0.056	0.020	(–0.095, –0.018)
			cardbill	689.443	912.893	(–1099.827, 2478.713)
		SL	mortality	–0.045	0.013	(–0.070, –0.019)
			cardbill	983.534	824.881	(–633.233, 2600.301)
	GSM	GLM	mortality	–0.058	0.023	(–0.104, –0.012)
			cardbill	275.740	991.980	(–1668.540, 2220.021)
		Tree	mortality	–0.044	0.013	(–0.071, –0.018)
			cardbill	844.911	823.143	(–768.448, 2458.271)
		GBM	mortality	–0.053	0.017	(–0.086, –0.020)
			cardbill	481.583	865.130	(–1214.071, 2177.237)
		SL	mortality	–0.047	0.012	(–0.070, –0.023)
			cardbill	640.864	819.759	(–965.863, 2247.591)
TMLE	Reg	GLM	mortality	–0.071	0.029	(–0.128, –0.013)
			cardbill	12.931	1226.643	(–2391.289, 2417.150)
		Tree	mortality	–0.043	0.015	(–0.072, –0.014)
			cardbill	1195.431	848.000	(–466.648, 2857.511)
		GBM	mortality	–0.061	0.021	(–0.102, –0.020)
			cardbill	546.337	994.494	(–1402.870, 2495.545)
		SL	mortality	–0.045	0.013	(–0.071, –0.019)
			cardbill	968.201	853.793	(–705.232, 2641.635)
	GSM	GLM	mortality	–0.064	0.025	(–0.112, –0.015)
			cardbill	612.648	936.713	(–1223.310, 2448.606)
		Tree	mortality	–0.043	0.013	(–0.069, –0.017)
			cardbill	1076.901	810.527	(–511.731, 2665.534)
		GBM	mortality	–0.056	0.017	(–0.090, –0.023)
			cardbill	745.969	837.604	(–895.734, 2387.673)
		SL	mortality	–0.046	0.012	(–0.070, –0.022)
			cardbill	891.970	811.203	(–697.988, 2481.927)
CF			mortality	–0.040	0.001	(–0.042 , –0.038)
			cardbill	993.260	60.931	(–873.836, 1112.684)

appear to well overlap in all cases. As in Li and Shen (2020), we can use Cochran's 0.25 rule to measure how balanced the variables are for each group. The measurements for *stent*, *acutemi* and *ves1proc* are greater than 0.25 in standardized differences, but the balance in these groups can be achieved through the propensity score adjustment. Several causal effect estimates are considered to

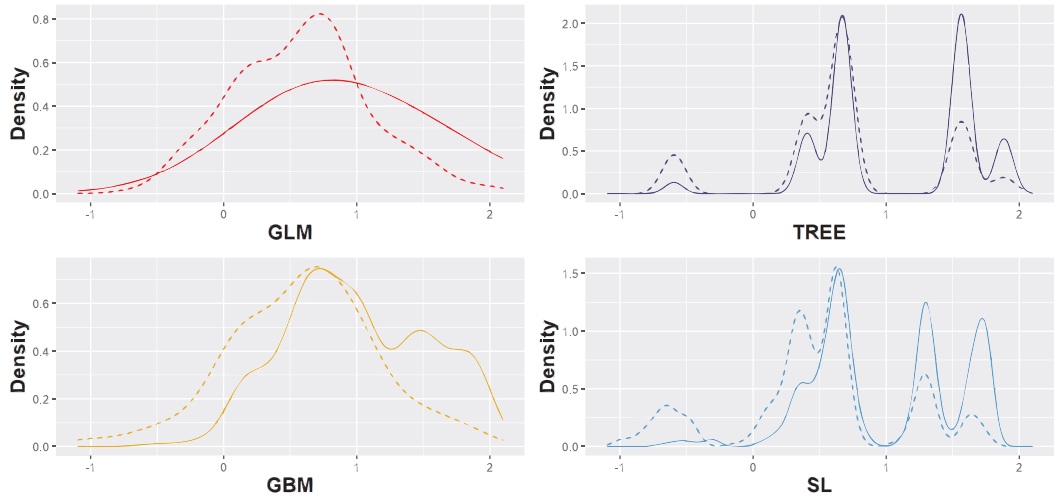

Figure 2: *Propensity score distributions for treatments with logistic, Tree, GBM, and SL algorithms.The Wald 95% CI is based on the robust SE. Each solid line and dotted line mean a group that has been treated and a group that has not been treated. Except for GLM model, machine learning models (TREE, GBM, and SL) show that the two graphs overlap quite well.*

compare the causal inference effects between *abcix* and normal treatment. Table 4 summarizes the causal effects of treatment on 996 patients for two response variables, *mortality* and *cardbill*. In AIPWE-SL and TMLE-SL, we used a SL model for estimating the propensity scores by optimally mixing five machine base learners, (i) *ipredbagg*, (ii) *ksvm*, (iii) *polymars*, (iv) *rpartPrune*, and (v) *KernelKnn*.

The results in Table 4 imply similar conclusions among all causal estimators (except for the regression estimator): patients with *abcix* treatment generally survive more than those of other group and the cost of *abcix* is reasonably determined. Notice that Reg shows somewhat different results for *cardbill*, for which we suspect model misspecification. The results of IPWE are also less reliable due to large standard errors. The ACE estimates for *mortality*, when the general treatment (0) and *abcix* treatment (1) are compared after six months, is $\hat{\tau} = -0.051$ with AIPWE-Reg-SL and $\hat{\tau} = -0.049$ with TMLE-Reg-SL, and both are significantly lower than 0. The mortality rate of *abcix* treatment is also lower than that of general treatment alone, demonstrating that treatment of *abcix* is much more effective. In respect of the cost, the ACE estimate of AIPWE-Reg-SL is $\hat{\tau} = \$983.534$ and that of TMLE-Reg-SL is $\hat{\tau} = \$968.201$. The cost of *abcix* treatment would be a little higher than that of general treatment alone, but their difference is not statistically significant after considering confounding effects. In summary, our results suggest that patients treated with *abcix* lived longer while requiring a little higher treatment costs.

5. Discussion

Doubly robustness (DR) is an important component in causal inference to correctly infer a target estimand after adjusting for potential confounders (Tsiatis, 2007). In observational studies, DR methods play an important role in estimating causal estimates if one of two working models (Q and π) is subject to model misspecification, but they may not work well if both components were not correctly specified. In this paper, we explore many DR inferential methods by replacing naïve linear regressions

with flexible ensemble algorithms that are commonly used in causal analysis. For the outcome model, we mainly focused on GSM (Helwig, 2020), which is much more flexible than a simple regression model or generalized additive model (GAM). To handle the propensity model, we applied several ensemble algorithms, such as decision tree, GBM, and SL algorithms, and demonstrate their usefulness under model misspecification, but of course, other machine learning algorithms could be explored to effectively reduce potential bias. CF is a newly introduced causal method that is based on random forest. Unlike conventional causal inference methods, CF does not require a complex model selection procedure and is free of model misspecification. However, we found that when working models are close to true models, DR estimators, such as TMLE and AIPWE, often outperform CF.

In this paper, we tried to compare several methods as much as possible under different simulation configurations, but our results are limited in that we cannot cover all possible scenarios. Therefore, care should be taken not to overgeneralize our findings. Main implication from this paper is that DR estimators, such as AIPWE and TMLE, with support from machine learning algorithms work well even when both models are not perfect. Their performance seem quite similar and it is hard to say that one is superior to the other in our setting. CF works very reliably and shows relatively more consistenet results when there is a highly nonlinear pattern in the data. However, these causal estimators cannot completely remove biases especially when informative covariates are omitted in the model. Hence, investigators should more focus on study design to get more information about the target variable as much as possible. Finally, in our experiments, it is assumed that the outcome is completely observed. However, the outcome is often masked due to censoring in many observational studies. It is interesting to explore causal machine learning algorithms and measure their performance in survival analysis (Choi *et al.*, 2021).

References

Athey S and Imbens G (2016). Recursive partitioning for heterogeneous causal effects. In *Proceedings of the National Academy of Sciences*, **113**, 7353–7360.

Athey S, Tibshirani J, and Wager S (2019). Generalized random forest, *The Annals of Statistics*, **47**, 1148–1178.

Austin PC (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research*, **46**, 399–424.

Breiman L (2001). Random forest, *Machine Learning*, **45**, 5–32.

Choi S, Choi T, Lee HY, Han SW, and Bandyopadhyay D (2021). Double-robust inferences for difference in restricted mean lifetimes using pseudo-observations, *Submitted*.

Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, and Davidian M (2011). Doubly robust estimation of causal effects, *The American Journal of Epidemiology*, **173**, 761–767.

Gruber S, and van der Laan MJ (2010). A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome, *The International Journal of Biostatistics*, **6**, 1557–4679.

Gulen H, Jens C, and Page TB (2020). An Application of Causal Forest in Corporate Finance: How-Does Financing Affect Investment?, *Microeconomics: Intertemporal Firm Choice & Growth*.

Helwig NE (2020). *Multiple and Generalized Nonparametric Regression*, SAGE Publications Limited.

Hernán MA and Robins JM (2020). *Causal Inference: What If*, CRC Boca Raton, Florida.

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

Lee BK, Lessler J, and Stuart EA (2010). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, *Statistical Science*, **22**, 523–539.

Li X and Shen C (2020). Doubly robust estimation of causal effect: upping the odds of getting the right answers, *Circulation: Cardiovascular Quality and Outcomes*, **13**, e006065.

McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, and Burgette LF (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models, *Statistics in Medicine*, **32**, 3388–3414.

Robins JM, Rotnitzky A, and Zhao LP (1994). Estimation of regression coeffcients when some regressors are not always observed, *Journal of the American statistical Association*, **89**, 846–866.

Robins JM, Sued M, Lei GQ, and Rotnitzky A (2007). Comment: Performance of double-robust estimators when inverse probability weights are highly variable, *Statistical Science*, **22**, 544–559.

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

Rubin D (1978). Bayesian inference for causal effects: The role of randomization, *The Annals of Statistics*, **6**, 34–58.

Schuler MS and Rose S (2017). Targeted maximum likelihood estimation for causal inference in observational studies, *American Journal of Epidemiology*, **185**, 65–73.

Stuart EA (2010). Matching methods for causal inference: A review and a look forward, *Statistical Science*, **25**, 1–21.

Tsiatis A (2007). *Semiparametric Theory and Missing Data*, Springer, New York.

Van der Laan MJ, Polley EC, and Hubbard AE (2007). Super learner, *Statistical Applications in Genetics and Molecular Biology*, **6**, 1544–6115.

Van der Laan MJ and Rose S (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer, California.

Wager S and Athey S (2018). Estimation and inference of heterogeneous treatment effects using random forest, *Journal of the American Statistical Association*, **113**, 1228–1242.

Received August 14, 2021; Revised November 17, 2021; Accepted November 25, 2021