

RESEARCH ARTICLE

Estimating treatment effects with machine learning

K. John McConnell PhD  | Stephan Lindner PhD 

Center for Health Systems
Effectiveness, Oregon Health & Science
University, Portland, Oregon

Correspondence

K. John McConnell, PhD, Center for Health
Systems Effectiveness, Oregon Health &
Science University, 3181 SW Sam Jackson
Park Rd., Mail Code MDYCHSE, Portland,
OR 97239.
Email: mcconnjo@ohsu.edu

Funding information

Silver Family Foundation (Portland, Oregon)

Abstract

Objective: To demonstrate the performance of methodologies that include machine learning (ML) algorithms to estimate average treatment effects under the assumption of exogeneity (selection on observables).

Data Sources: Simulated data and observational data on hospitalized adults.

Study Design: We assessed the performance of several ML-based estimators, including Targeted Maximum Likelihood Estimation, Bayesian Additive Regression Trees, Causal Random Forests, Double Machine Learning, and Bayesian Causal Forests, applying these methods to simulated data as well as data on the effects of right heart catheterization.

Principal Findings: In Monte Carlo studies, ML-based estimators generated estimates with smaller bias than traditional regression approaches, demonstrating substantial (69 percent–98 percent) bias reduction in some scenarios. Bayesian Causal Forests and Double Machine Learning were top performers, although all were sensitive to high dimensional (>150) sets of covariates.

Conclusions: ML-based methods are promising methods for estimating treatment effects, allowing for the inclusion of many covariates and automating the search for nonlinearities and interactions among variables. We provide guidance and sample code for researchers interested in implementing these tools in their own empirical work.

KEYWORDS

machine learning, observational research, treatment effects

1 | INTRODUCTION

Machine learning (ML) applications are becoming increasingly common in health services research.^{1–3} To date, most of this research has focused on problems of prediction, but recent developments in ML expand their applications beyond predictive models into the realm of statistical inference, creating opportunities for broader use in the field.

The purpose of this article is to highlight methods that allow for ML to be used for estimation and inference and to compare them to traditional regression-based approaches when estimating average treatment effects in a cross-sectional setting. We begin by providing a brief (nonexhaustive) introduction to ML

methodologies. Next, we introduce methods that allow for the use of ML in estimating average treatment effects under conditions where confounding factors are observed, a situation often characterized as “unconfoundedness,” “exogeneity,” or “selection on observables.”⁴ We then assess the performance of these estimators through a variety of Monte Carlo simulations, focusing on cross-sectional environments where the relevant covariates are known and observed, but correlated with treatment and outcome. We use these simulations to highlight the various strengths of different estimators under different conditions. We also demonstrate the performance of these methods when applying them to data on the effect of right heart catheterization (RHC), first analyzed by Connors et al⁵ and widely reanalyzed.^{6–16} We provide

example R code to demonstrate how researchers can call these methods to use them in their own research.

2 | BACKGROUND

Machine learning (ML) encompasses a wide range of applications and methodologies; recent review articles may serve as useful introductions for those new to the field.¹⁷⁻²⁴ ML applications are typically divided into “supervised” and “unsupervised” branches, with “supervised” ML concerning the prediction of an outcome (Y) based on a set of covariates (X), an approach that has similarities to traditional regression models. Methodologies that fall under the realm of supervised learning include regularized regression (including Least Absolute Shrinkage and Selection Operator, or “LASSO,” ridge, and elastic net), regression trees, random forests, gradient boosting, and neural networks. “Ensemble” methods use multiple algorithms (eg, LASSO, random forests, and neural nets) to obtain better predictive performance than could be obtained from a lone algorithm.^{25,26}

In contrast to traditional statistical models, ML can largely be viewed as tools that excel in prediction. Recent studies in health services research have exploited this aspect, using ML for prediction to provide insights beyond what had been achieved with traditional statistical models.^{1,3,24,27,28}

More recently, ML approaches have been expanded beyond a focus of *prediction* into the area of *inference*, for example, estimating the effect of a variable on an outcome, with a corresponding confidence interval. In this article, we focus on estimating the average treatment effect (ATE) of a binary treatment variable (D) on an outcome Y, assuming that D is not randomly assigned, but confounded by some vector of covariates (X) which also influence Y. Under the assumption of exogeneity (or “unconfoundedness”),²⁹ it is possible to estimate the effect of D on Y by controlling for the relevant X variables and the ways that X influences D and Y. This is a common research design in health services, where investigators with observational data wish to isolate the effect of some treatment (eg, a drug, intervention, or delivery model) on health outcomes, utilization, or expenditures.

In this setting, estimates of the average treatment effect will be biased if relevant X variables are not observed or included or if the way in which they influence D or Y is not correctly specified. The proliferation of large datasets with thousands of variables offers new opportunities to control for confounding, but also creates new challenges in terms of how to do so. In the past, researchers have traditionally succumbed to ad hoc decisions, perhaps choosing a parsimonious model over one that includes hundreds of variables. In addition, researchers have often assumed that the relationship between Y and X and D and X could be captured via a linear functional form (perhaps with some interactions among key variables). These are imperfect solutions. In contrast, ML approaches can automate many of these choices, algorithmically selecting the relevant variables and functional forms (even in cases where the number of available X variables (p) is greater than the number of observations (N)),

generating estimates that reduce the bias that would be present in more traditional parametric approaches, and potentially producing unbiased estimates.

We highlight six ML-based methods that can be used to produce estimates of average treatment effects under conditions of exogeneity. In rough order of chronological appearance in the literature, these include (a) Targeted Maximum Likelihood Estimation (TMLE),^{30,31} (b) Bayesian Additive Regression Trees (BART),³² (c) Causal Random Forests (CRF),^{33,34} (d) Double Machine Learning (DML),³⁵ and (e) Bayesian Causal Forests (BCF).³⁶ We also assess a variation of BART, (f) ps-BART, which incorporates a propensity score modeling step in the estimation, and can be seen as a bridge between BART and BCF. We describe each of these in brief below.

TMLE is a well-established method that has been described in a variety of books, articles, and tutorials.^{30,31,37,38} TMLE is a double-robust, two-stage estimator. The first stage produces an initial estimate of $Q^0 = E(Y|D, X)$ and the propensity score, $P(D = 1|X)$. The propensity score is used to construct a variable referred to as the “clever covariate,” C, defined as $(1/(P(D = 1|X))$ for individuals in the treatment group and $(-1/(P(D = 0|X)))$ for individuals in the comparison group. In the second stage, Y is regressed (eg, via generalized linear model) on C and Q^0 , with the coefficient on Q^0 is constrained to 1; this model iteratively generates an empirical density function for Y which is then used to calculate the average treatment effect. Because $E(Y|D, X)$ and $P(D|X)$ can be conceptualized as prediction problems, ML can be incorporated in the TMLE procedure to allow for semiparametric estimation, including ensemble approaches performed via the “Super Learner” method, which combines ML algorithms through weighting to produce an estimate that minimizes the cross-validated mean squared error.³⁹ Standard errors are calculated using the influence curve method.^{40,41}

The Bayesian Additive Regression Trees (BART) estimation procedure uses many single “regression trees” to generate a predictive model. Regression trees are an early type of machine learning (dating back to the 1970s^{42,43}) that recursively partitions a dataset into smaller groups and then fits a simple model (constant) for each subgroup. The final partitioning can be represented visually as a decision tree. Generally, single trees tend to be unstable and relatively poor predictors. BART uses regression trees as a starting point and combines two elements: a sum-of-trees model and a regularization prior.⁴⁴ The sum-of-trees model can be conceptualized as using a tree to estimate $E(Y|D, X)$, taking the residuals, and subsequently fitting additional trees on the residuals, with a sum-of-trees approach that has flexibility and is adept at capturing both nonlinearities and interactions that generate good predictive capabilities. The regularization (or “shrinkage”) priors include parameters that allow each tree to contribute only a small part to the overall fit. If BART produces highly accurate predictions, it can be used to generate estimates of the average treatment effect taking differences in the predicted values of $E(Y|D = 1, X) - E(Y|D = 0, X)$.³² BART is a Bayesian specification with estimation producing a posterior distribution.

Bayesian Causal Forests (BCF) take the BART approach and extend it to improve its performance in estimating treatment effects.³⁶

The BCF model starts with the BART approach and makes two modifications. The first modification is based on an observation that BART, when used to estimate treatment effects, may exhibit bias, due in part to the regularization priors.⁴⁵ BCF corrects for this bias by including an estimate of the propensity score in the group of covariates passed to the BART model. (This simple modification is referred to as ps-BART). Second, like BART, BCF estimates the conditional mean based on the X covariates (and propensity score), but also includes a function to estimate the treatment effect directly. Like BART, estimation with ps-BART or BCF produces a posterior distribution.

Generalized Random Forests^{33,34} build on the notion of random forests, a machine learning approach introduced by Breiman.⁴⁶ Random forests start with the regression tree approach, but rather than using one tree, build multiple trees (each faced with a random subset of features and data, to increase diversity in the trees). The final prediction is based on an average of the individual decision tree estimates. The GRF method adapts the random forest approach to produce estimates of treatment effects. The GRF approach can be conceptualized as changing the outcome, using the sample average treatment effect ($Y(D = 1) - Y(D = 0)$) in the end node, or “leaf”, as opposed to the traditional outcome (Y). Whereas a typical regression tree is optimized to identify variables and values that minimize mean squared error, GRF trees are designed to identify variables and values where the treatment effects differ most. The overall algorithm then mimics the setup of a standard random forest, with the estimates of the treatment effect based on an average of multiple trees. Athey and colleagues show it is asymptotically normal.³³

Double Machine Learning³⁵ (DML) is based on a simple and intuitive insight initially presented by Frisch and Waugh in 1933 and commonly referred to as the Frisch-Waugh-Lovell theorem.^{47,48} A result of this theorem is that the estimate of the parameter β_D in the equation $Y = \beta_0 + D^* \beta_D + \theta X + e$ can be obtained by (a) regressing Y on X (leaving D out) and saving the residuals (\hat{e}); (b) regressing D on X and saving the residuals ($\hat{\omega}$); and (c) regressing \hat{e} on $\hat{\omega}$ (“residual-on-residual”). The coefficient on $\hat{\omega}$ in this “residual on residual” regression is asymptotically (and algebraically) equivalent to the estimate of β_D . Recognizing that obtaining the residuals \hat{e} and $\hat{\omega}$ can be reduced to a prediction problem ($E(Y|X)$ and $E(D|X)$) opens the door for semiparametric ML algorithms to produce those residuals.

Chernozhukov and colleagues note that, when the functional forms are complex and when semiparametric ML techniques are used to generate residuals, a simple residual-on-residual regression will be asymptotically biased. Their solution to this problem is to use “sample splitting” or “ k -fold cross-fitting”, an approach that builds on work by Angrist and Krueger⁴⁹ and has analogs in the cross-validation routines that are standard in ML algorithms. In the k -fold cross-fitting approach, a dataset of sample size N is partitioned into K equal parts. Construction of the residuals is based by first estimating the models to predict $E(Y|X)$ and $E(D|X)$ on a subset of data that excludes the K th partition, and then using those parameters to construct predicted estimates and residuals into the K th partition. The residual-on-residual regression is conducted within the K th

partition, and the final estimate of the ATE is based on the average of K estimates conducted for each partition. At a minimum, the researcher should choose K to have a value of 2; higher folds are possible but come with increased computational burden.

This sample splitting approach is asymptotically efficient but creates an additional source of uncertainty in finite samples. Thus, Chernozhukov and colleagues propose that the DML routine be repeated S times, with S taking on a larger number (20, 50, 100) as a mechanism for capturing additional variation introduced via sample splitting.

These six methods (TMLE, BART, ps-BART, BCF, GRF, and DML) provide new opportunities for researchers interested in estimating average treatment effects where there are potentially a large set of candidate explanatory variables or where the functional forms are unknown and possibly complicated. However, a greater appreciation for the performance of these estimators and the choices available to the researcher may be necessary to increase their use in the field of health services research. To that end, we turn to the practical questions of how these estimators can be deployed and how they perform.

3 | METHODS

To further explore these methods, we conduct a variety of Monte Carlo simulations, using a data generating process (DGP) that includes complex, nonlinear relationships between Y and X and D and X . This DGP is easily modified to increase both the extent to which there is confounding with the observed X variables, as well as the number of X variables under consideration (ie, the width of the database). Our DGP does not necessarily reflect any specific scenarios that would arise in health services, but reflects a flexible structure that mimics empirical problems that could be challenging for the traditional statistical toolbox: for example, when the set of candidate X variables is large and when the functional forms are complex. Furthermore, we note that this DGP is consistent with the assumptions of “unconfoundedness” or “exogeneity.”

Our DGP is based on the following model:

1. $Y_i = D_i^* \beta_D + \gamma^* \cos^2(x_i' \beta) + u_i$
2. $D_i \sim \text{Bernoulli}(p_i)$, and $\text{logit}(p_i) = \sin(x_i' \beta) + \cos(x_i' \beta) - \alpha$

The treatment effect, β_D , is set at $\beta_D = 1$; $u_i \sim \text{Normal}(0,1)$, and $\alpha = 0.5$, so that approximately 50 percent of subjects receive treatment. The variable x_i represents a vector of J covariates, generated from a multivariate normal distribution; β is a vector of J parameters with $\beta_j = 1/j$ (for $j = 1, \dots, J$).^a (In other words, as we increase J , we add covariates, but each covariate is slightly less correlated with the treatment and outcome). The parameter γ is a scalar whose role is described below. Treatment (D_i) is a binary variable governed by p_i , the probability of treatment.

We can adjust parameters in this DGP to create more challenging estimation problems. In particular, we focus on J (the number of X variables) and γ , a parameter that can be changed to increase or

TABLE 1 Simulation details

DGP	Details		Key parameters
Data generating processes			
DGP	<ul style="list-style-type: none">Y is a function of D and a complex functional of X variables ($\gamma \cdot \cos^2(x_i' \beta)$) drawn from a multinomial normal distributionD is binomially distributed according to $\text{logit}(p_i) = \sin(x_i' \beta) + \cos(x_i' \beta) - \alpha$“True” treatment effect has a value of 1.0. (See text for additional details)		<ul style="list-style-type: none">γ, the confounding weightN, sample sizeJ, the number of correlated covariatesP, the number of uncorrelated (noise) covariates
Estimators	Details	Implementation in R	Notes
Estimators			
OLS	Linear regression of Y on D, X	lm command	
AIPW	Augmented Inverse Probability Weighted estimator	CausalGAM package	
TMLE	Targeted Maximum Likelihood Estimation	tmle package	Uses random forest ML algorithm as underlying estimator Other Super Learner library options are possible
BART	Bayesian Additive Regression Trees	bart_est package	Other R packages available; new BART package allows for parallel processing Standard errors estimated from posterior distribution
ps-BART	Bayesian Additive Regression Trees, with propensity score included as covariate	bart_est package Propensity score generated with logistic regression	
BCF	Bayesian Causal Forest	bcf package	
GRF	Generalized Random Forests	grf package	
DML	Double machine learning algorithm	Code adapted from R code provided by Chernozhukov et al 2017, available at https://www.aeaweb.org/articles?xml:id=10.1257/aer.p20171038	Uses random forest ML algorithm as underlying estimator k = 2 for k-fold cross-fitting Procedure repeated over S = 20 iterations

decrease the confounding between X, Y, and D. For example, with $\gamma = 0$, there is no confounding by X and OLS will produce unbiased estimates of β_D . As we move γ away from zero, Y is affected more strongly by the X variables, which also influence the probability of treatment (eg, confounding increases), and we anticipate greater bias in our estimates of the treatment effect.^b We can also include additional noise variables (P)—randomly generated covariates with no correlation to Y or D—to understand whether their inclusion reduces the effectiveness of our estimators.

Our "baseline" DGP is defined as $\gamma = 2$; $J = 10$; $N = 5000$; and $P = 0$. We then describe performance as we systematically vary these parameters, while holding others constant:

1. Change in confounding strength, varying γ (0, 1, 2, 3)
2. Change in sample size, varying N (1000, 2000, 5000, 10 000)
3. Change in number of variables correlated with Y, varying J (10, 20, 50, 100)
4. Change in the number of "noise" variables, varying P (0, 10, 40, 90).

We conduct M = 1000 replications for each sample.

Our primary simulations include two traditional approaches: OLS and an augmented inverse propensity weighted (AIPW) estimator,

a "doubly robust" estimator that estimates the propensity score of treatment and incorporates this information to conduct a weighted OLS regression.⁴ Our OLS and AIPW models include all X variables in an additive linear form (ie, the functional form is incorrectly specified).

These methods are compared to six ML-based approaches: TMLE, BART, ps-BART, BCF, GRF, and DML (with DML implemented with 2-fold sample splitting and 20 iterations). The TMLE and DML methodologies require us to specify the underlying ML algorithm; we use random forests. Random forests represent a popular ML approach that has generally been shown to perform well in prediction problems without a need for extensive tuning.⁵⁰

We conduct 1000 replications for each sample. For each estimator, we calculate the bias (as a percent deviation from the true estimate), root mean squared error (RMSE), median absolute deviation (MAD), the average estimated standard error for each estimate, across all simulations, and the coverage rates for nominal 95% confidence intervals (ie, the percent of simulations for which the 95% confidence interval of our estimate includes the true treatment effect). Table 1 provides additional detail about our estimators and outlines our data generating process, which is designed to allow the researcher to arbitrarily dial up the amount of (observable) confounding or adjust the number of covariates to be considered.

In addition to our Monte Carlo analyses, we also provide an example of the performance of these measures when estimating the effects of right heart catheterization (RHC) on 30-day mortality, using observational data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). RHC is a procedure used for critically ill patients, which, historically, had been assumed to be beneficial for most patients. However, the SUPPORT trial found that, after adjusting for a range of covariates, patients who received RHC appeared to have higher mortality relative to those who did not receive RHC.⁵ Many researchers hypothesized that this unexpected result was driven by an inability to fully adjust for unobserved confounding. The SUPPORT data have been widely reanalyzed⁶⁻¹⁶ and are publicly available <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/rhc.csv>.

These data include 5735 individuals who were admitted to or transferred to an ICU in the first 24 hours after entering the study, with 2184 “treated” patients (receiving RHC within 24 hours of admission) and 3551 controls. The binary outcome variable is survival at 30 days postadmission. We also include 72 covariates. These data and covariates (as well as comparisons between treated and untreated populations) are described in detail in Hirano and Imbens.⁶

Our analyses of the SUPPORT data are designed to make two contributions. First, we provide R code that analyzes the SUPPORT data to produce estimates of the ATE and associated standard errors, which should allow other researchers to quickly adapt these calls to their own data structures. Second, we use the SUPPORT data to demonstrate their effectiveness in a real-world data, noting that in this particular case, the algorithms produce similar estimates and still indicate a positive association between RHC and mortality, suggesting that some unobserved confounding or endogeneity persists and is not resolved with the machine learning advancements.

3.1 | Limitations

There are important limitations to this study. First and foremost, our analyses represent estimates of average treatment effects under exogeneity. These conditions may be quite unusual in health services, where unobserved confounding and omitted variable bias present problems that are more likely to be resolved through other statistical approaches, including instrumental variables, regression discontinuity, or longitudinal studies. Our reanalysis of the SUPPORT data provides a useful comparison to previous studies, but the “true” effect of RHC is unknown, so the results must be viewed in this context.

The approaches described here can be seen as relying largely on intensive, computer-driven searches to pull out a single parameter. Despite the computing power inherent in these methods, they may be suboptimal to structural approaches or more refined conceptual models that may direct modeling decisions.

Our Monte Carlo simulations test a variety of DGPs, but this list is by no means exhaustive. There may be DGPs for which our general results do not hold, including DGPs that are similar to what we present but for which the sample sizes are larger than those we test. In addition, our outcome variable is normally distributed; we

do not assess performance for DGPs with alternative forms, such as highly skewed or non-negative distributions that are common in health services research. Furthermore, while we test for average treatment effects, we do not assess performance in estimating heterogeneous treatment effects. Some of the approaches presented here (especially GRF and BCF) may be particularly strong in these areas.

Two of our methods, TMLE and DML, require the analyst to specify the underlying ML algorithm. We use random forests, but do not compare the performance of a range of other algorithms (including the rich possibilities within the TMLE Super Learner ensemble). Finally, we note that our review does not include the universe of ML-based methods for estimating treatment effects; this is a fast-moving field, and we anticipate that researchers will continue to propose alternative methods, some of which may have better performance than those reviewed here.

4 | RESULTS

Table 2 displays results for our baseline simulation ($\gamma = 2$; $J = 10$; $N = 5000$; and $P = 0$). The bias for OLS is the largest (18.8 percent), and the doubly robust AIPW estimator does not fare much better. Turning to our ML-based approaches, ps-BART and BCF are the standouts in this scenario, exhibiting a bias close to zero, the smallest RMSE, and coverage rates at 94 percent and 95 percent. The next best performer is DML, followed by GRF, BART, and TMLE. TMLE has the largest bias (5.9 percent) and smallest coverage (24.6 percent) among the ML-based estimator. Nonetheless, all of the ML-based estimators exhibit significant improvements over AIPW, with reductions in bias ranging from 69 percent to 98 percent.

In order to understand the performance of different estimators under different conditions, we test performance across four scenarios, moving away from our baseline simulation ($\gamma = 2$; $J = 10$; $N = 5000$; and $P = 0$) by systematically changing sample size (N), confounding (γ), the number of variables correlated with Y (J), and number of noise variables (P). We provide full results (bias, RMSE, MAD, standard error, and coverage) for all estimators in Appendix S1 (Section 1). To provide a more cohesive understanding of performance, we graph bias for selected estimators (OLS, AIPW, TMLE, GRF, BCF, DML) under these four scenarios. To simplify our figure, we leave off two estimators—BART and ps-BART—because ps-BART dominates BART in all scenarios, and because results for BCF and ps-BART are close to identical (see Appendix S1; Section 1 for additional details). We display our results in Figure 1, graphing bias for each estimator as we vary the amount of confounding (γ , top left), observations (N , top right), number of variables correlated with Y (J , bottom left), and number of noise variables (P , bottom right). We begin by describing general patterns, and then drill down to specific instances of note.

Across the 16 different DGPs in Figure 1, ML-based approaches perform better than the OLS and AIPW methods. When the number of covariates is small, BCF is typically the top performer. Panel A displays performance as we increase the confounding weight. BCF is

TABLE 2 Simulation results for baseline scenario

Estimators	% bias	RMSE	MAD	Mean SE	95% CI coverage rate
OLS	18.8	0.194	0.282	0.036	0.8%
AIPW	18.3	0.188	0.274	0.037	0.5%
TMLE	5.9	0.069	0.087	0.017	24.6%
BART	5.7	0.070	0.086	0.035	60.2%
ps-BART	0.7	0.033	0.035	0.032	94.0%
BCF	0.4	0.033	0.035	0.032	95.0%
GRF	4.1	0.055	0.062	0.033	73.6%
DML	2.5	0.044	0.043	0.035	87.8%

Note: Simulation results based on 1000 replications.

Baseline scenario: γ (confounding weight) = 2; N (sample size) = 5000; J (correlated covariates) = 10; P (noise covariates) = 0.

TMLE and DML take random forests as underlying ML algorithm. DML run with 2-fold sample splitting and 20 iterations.

Abbreviations: AIPW, Augmented Inverse Probability Weighting; BART, Bayesian Additive Regression Trees; BCF, Bayesian Causal Forest; DML, Double Machine Learning; GRF, Generalized Random Forest; MAD, mean absolute deviation; OLS, ordinary least squares; ps-BART, Bayesian Additive Regression Trees with propensity score; RMSE, root mean square error; TMLE, Targeted Maximum Likelihood Estimation.

the top performer, followed by GRF, DML, and TMLE. Remarkably, BCF is essentially unbiased, even as confounding increases substantially (eg, 38 percent bias for OLS when $\gamma = 4$ compared to a 1.0 percent bias for BCF).

The top right panel (B) displays changes in bias with increasing sample size. Among the ML estimators, increased sample size tends to improve performance, in contrast to OLS and AIPW.

Panels C and D display bias as we increase the number of covariates. Across all ML approaches, increasing the number of covariates increases the bias, although the performance of BCF and GRF degrades more substantially. With $J = 200$, the top performer is the DML estimator. Adding noise variables generates similar performance changes, with DML exhibiting the lowest bias among all estimators.

We provide complete output (including RMSE, MAD, and coverage rates) for each of these simulations in the Appendix S1 (Table S1). We also provide results from a variety of other tests that assess changes in the performance in TMLE and DML that occurs with different underlying ML approaches (Table S2), and in changing some parameters in the DML approach (Table S2).

Table 3 displays our results from analyses of the SUPPORT data, including information on computing time required to produce treatment effects for this dataset, which includes 5735 observations and 72 covariates. The OLS and AIPW estimators have similar point estimates of the effect of right heart catheterization 6.2 percent, although the AIPW measure has a higher standard error. With the exception of the TMLE estimator using random forests, the ML-based point estimates are generally lower than OLS and AIPW estimates (ranging from 3.9 percent to 5.1 percent), but none suggest a protective or null effect of RHC.

The most striking exception to these estimates are those produced by TMLE. We show two estimates for TMLE: one using a

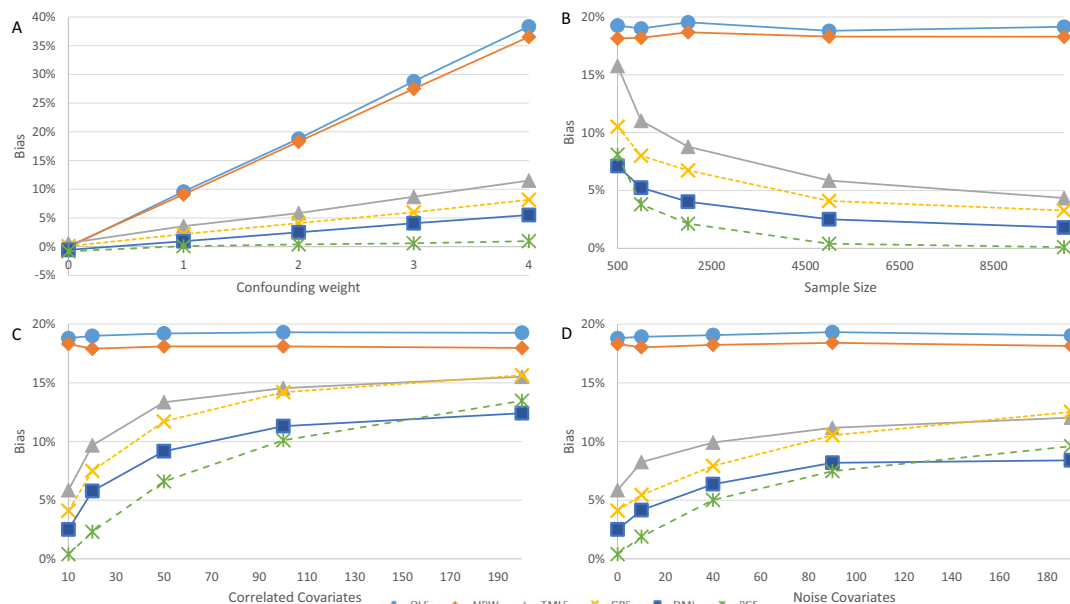


FIGURE 1 Bias of different estimators, under different DGP scenarios. A, Changes in confounding. B, Change in sample size. C, Change in correlated covariates. D, Change in noise variables. Where not indicated, γ (confounding weight) = 2; N (sample size) = 5000; J (correlated covariates) = 10; P (noise covariates) = 0. AIPW, Augmented Inverse Probability Weighting; BCF, Bayesian Causal Forest; DML, Double Machine Learning; GRF, Generalized Random Forest; OLS, ordinary least squares; TMLE, Targeted Maximum Likelihood Estimation. TMLE and DML take random forest as the underlying ML algorithm. DML run with 2-fold sample splitting and 20 iterations. [Color figure can be viewed at wileyonlinelibrary.com]

Note: BART and ps-BART are not included because BCF dominates BART in all simulations, and results for ps-BART and BCF are almost identical

TABLE 3 Estimates of effect of RHC (standard errors in parentheses) on 30-day mortality

Method/source	Estimate of treatment effect on 30-d mortality	Computing time
Unadjusted	7.4% (1.3)	<1s
OLS	6.2% (1.3)	<1s
AIPW	6.2% (1.6)	0:30
TMLE (Random Forest)	7.0% (0.3)	6:37
TMLE (Super Learner)	4.7% (0.6)	7:40
BART	5.1% (1.3)	1:18
ps-BART	5.0% (1.2)	1:18
BCF	4.5% (1.3)	5:32
GRF	4.3% (1.3)	1:40
DML	3.9% (1.3)	0:38

Note: Estimates based on data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT), a study of the effectiveness of right heart catheterization (RHC). Data include 5735 individuals who were admitted to or transferred to an ICU in the first 24 h after entering the study, with 2184 “treated” patients (receiving RHC within 24 h of admission) and 3551 controls. Adjusted estimates are based on inclusion of 72 covariates that capture patient health risk and demographics. For comparison, we present estimates provided by a previous analysis of the SUPPORT data by Hirano and Imbens, who use propensity score weighting and regression adjustment. Standard errors are presented in parentheses. Time is presented as minutes:seconds. Run times are based on computations on a Dell PowerEdge R730 Server with dual Intel Xeon CPU E5-2640 v3 processors with a clock speed of 2.60 GHz. TMLE and DML take random forest as the underlying ML algorithm. DML run with 2-fold sample splitting and 20 iterations.

random forest as the underlying algorithm (in line with our simulations) and one using Super Learner to include an expanded ensemble set of estimators. The TMLE estimate using only a random forest produces an estimate of 7.0 percent—a figure that would seem to be out of line with other estimates, lying somewhere between the adjusted and unadjusted OLS estimates, with a standard error that is approximately 75 percent too small. The TMLE estimate that incorporates Super Learner generates a more reasonable estimate (4.7 percent), although the estimated standard error still appears too small.

Table 3 also provides an indication of computing time, with GRF and DML being relatively quick, and BCF and TMLE being the most computationally intensive (taking approximately five to seven minutes in this example). The slower performance for TMLE in the SUPPORT application was not replicated in our simulations; in fact, TMLE was typically among the fastest of all estimators when applied to our simulation data.

In our Appendix S2, we provide the sample R code for using each of these estimators, using the SUPPORT data as an example test case.

5 | DISCUSSION

In this study, we compared 6 ML-based estimation approaches to more traditional, parametric approaches to estimating treatment effects. Under conditions of exogeneity, our ML methods estimators routinely outperformed traditional methods, often with substantial reductions in bias. However, the results of our simulation and reanalysis of the SUPPORT data also show that ML estimators do not always eliminate bias and are not a panacea for the problems of unobserved confounding. We offer the following guidance for researchers.

1. Our results suggest there is no single best ML-based estimator, although the ps-BART and BCF estimators are strong contenders for a range of scenarios we examine. BCF is easy to implement and requires relatively little tuning or adjustment of hyperparameters from the researcher, although it is more computationally intensive. In our simulations, the stand-alone BART approach appears to offer no advantages over ps-BART.
2. Researchers with a large set of covariates should be cognizant of the potential sensitivity of estimators to the inclusion of additional variables. In the specifications we tested, DML was the top performer when the number of X variables exceeded 150.
3. For ease of use and computational efficiency, TMLE may be an attractive option. However, performance may depend heavily on the underlying algorithm, as evidenced by the differing estimates produced in the SUPPORT example and shown in the Appendix S2. The GRF and BCF algorithms are similarly easy to implement.
4. Researchers specifically interested in heterogeneous treatment effects may want to consider the BCF and GRF estimators. Although the focus of this paper was on average treatment effects, the design of BCF and GRF may afford them advantages when effects are heterogeneous.

Table 4 provides a summary of guidelines and considerations for the practitioner.

Our analyses were conducted with default ML settings. However, the performance of ML-based algorithms may be further improved by additional tuning of specific hyperparameters (eg, the number of trees to grow in a random forest algorithm). Tuning is performed by a process of selecting parameters that maximize the predictive accuracy of the model in question. The process of tuning an algorithm may be conceptually novel to many health services researchers; Einav et al¹ provide an example of a clear and careful exposition of a tuning process in their appendix.

These simulations are conducted under the assumption of exogeneity; many studies of health services research will not withstand this assumption and will face omitted variables or unobserved confounding. Application of these methods under those conditions will not ameliorate the fundamental problems of identification. Nonetheless, we anticipate that ML-based methods will eventually make contributions in these areas. For example, several studies have explored ML-based approaches in the first stage of

TABLE 4 Comparison of ML-based approaches

	TMLE	BART and ps-BART	BCF	GRF	DML
Key references	van der Laan and Rose 2011; Schuler and Rose 2017	Hill 2011 (BART); Hahn, Murray and Carvalho (ps-BART)	Hahn, Murray, Carvalho 2018	Athey, Tibshirani, and Wager 2019; Wager and Athey 2018	Chernozhukov et al 2017
Performance (bias)	Reductions in bias typically not as large as ps-BART, BCF, GRF, or DML	ps-BART consistently outperforms BART and is often a top performer Sensitive to higher-dimensional data	Low bias with small number of covariates Sensitive to higher-dimensional data	Generally slightly worse than BCF, ps-BART, or DML; sensitive to higher-dimensional data	Generally slightly worse than BCF or ps-BART; smaller drop in performance with higher-dimensional data
Performance (standard errors)	Standard errors provided via "influence curve" are sensitive to underlying ML/Super Learner choices and may be overly conservative in some cases Bootstrapping may provide more reliable standard errors and confidence intervals	Produces posterior distribution; standard error can be computed manually	Produces posterior distribution; standard error can be computed manually	Constructed via the delta method	Standard errors estimated via averaging across iterations and split-sample folds (formulae provided in Chernozhukov et al 2017)
Ease of use	Very easy; one to two lines of code; documentation and package available in R. Generally fast, although speed may depend on underlying ML algorithms and structure of data	Some coding required to produce treatment effects for continuous or dichotomous outcome and to convert posterior distribution to parameter estimate and standard error	Straightforward to implement; minimal coding needed to convert posterior distribution to parameter estimate and standard error	Very easy; one to two lines of code; documentation and package available in R	Some coding in R required; code provided in appendix material by Chernozhukov et al 2017
Additional considerations	A variety of Super Learner ensemble library options are available; eg, recent developments support inclusion of "Highly Adaptive Lasso (HAL)" estimator	"Overfitting" propensity score may lead to unstable estimates; recommend may be safer to generate propensity score through simple logistic estimation	"Overfitting" propensity score may lead to unstable estimates; recommend may be safer to generate propensity score through simple logistic estimation May be particularly useful in assessing heterogeneous treatment effects	May be particularly useful in assessing heterogeneous treatment effects	Programmer must choose number of split-sample folds (K; eg, 2, 5, 10) and iterations (S) (eg, 20, 50, 100). Larger values of K or S may reduce bias but increase computational burden DML can incorporate ensemble methods and an "interactive" estimation mechanism, both of which may further reduce bias
Theoretical support for consistency and asymptotic normality?	Yes ⁵⁷	For BART, no; for ps-BART, theoretical support for consistency in linear and partially linear semiparametric context has been provided ^{45,58}	Theoretical support for consistency in linear and partially linear semiparametric context has been provided ^{45,58}	Yes ³³	Yes ⁵⁹

instrumental-variable models,⁵¹⁻⁵⁴ while recent working papers focused on difference-in-difference models have suggested a role for “debiasing” approach of DML.^{55,56}

6 | CONCLUSIONS

Machine learning's importance in health services research is likely to grow substantially in the near future. Applications that allow for statistical inference will be particularly important. Our results suggest that, when estimating treatment effects under exogeneity, these methods can reduce bias relative to more traditional approaches even when observed confounding is considerable. In some cases, the reduction in bias is substantial. We provide guidance for researchers interested in deploying these models in their ongoing empirical work.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported in part by a grant from The Silver Family Foundation (Portland, Oregon). We are grateful to Thomas Meath, Matt Georg, and Kyle Tracy for help with coding, statistical analyses, and computing power considerations.

CONFLICTS OF INTEREST

We have no conflicts of interest and no other disclosures.

ORCID

K. John McConnell  <https://orcid.org/0000-0003-1280-3196>

Stephan Lindner  <https://orcid.org/0000-0002-9577-6287>

ENDNOTES

^a The basis for this data generating process was proposed by Gabriel Vasconcelos as a blog post on the InsightR website. <https://insightr.wordpress.com/2017/06/28/cross-fitting-double-machine-learning-estimator/> [Accessed June 2, 2018].

^b If $Y = \alpha + \beta D + \theta X + \varepsilon$, then the OLS estimate of β is

$$\hat{\beta} = \beta + \theta \frac{\text{cov}(D, X)}{\text{var}(D)} + \frac{\text{cov}(D, \varepsilon)}{\text{var}(D)}$$

The last term will go to zero as N gets large. If D and X are correlated and θ is not zero, then the estimate of β will be biased.

REFERENCES

- Einav L, Finkelstein A, Mullainathan S, Obermeyer Z. Predictive modeling of U.S. health care spending in late life. *Science*. 2018;360(6396):1462-1465.
- Rose S. Robust machine learning variable importance analyses of medical conditions for health care spending. *Health Serv Res*. 2018;53(5):3836-3854.
- Rose S, Bergquist SL, Layton TJ. Computational health economics for identification of unprofitable health care enrollees. *Biostatistics*. 2017;18(4):682-694.
- Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1):4-29.
- Connors AF, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*. 1996;276(11):889-897.
- Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Method*. 2001;2(3):259-278.
- Pingel R. Estimating the variance of a propensity score matching estimator for the average treatment effect. *Observ Stud*. 2018;4:71-96.
- Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2018;113(521):390-400.
- Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199.
- Traskin M, Small DS. Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Stat Biosci*. 2011;3(1):94-118.
- Altonji JG, Elder TE, Taber CR. Using selection on observed variables to assess bias from unobservables when evaluating Swan-Ganz catheterization. *Am Econ Rev*. 2008;98(2):345-350.
- Bhattacharya J, Shaikh AM, Vytlacil E. Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization. *Am Econ Rev*. 2008;98(2):351-356.
- Li Q, Racine JS, Wooldridge JM. Estimating average treatment effects with continuous and discrete covariates: the case of Swan-Ganz catheterization. *Am Econ Rev*. 2008;98(2):357-362.
- Bhattacharya J, Shaikh AM, Vytlacil E. Treatment effect bounds: an application to Swan-Ganz catheterization. *J Econometr*. 2012;168(2):223-243.
- Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998;54(3):948-963.
- Rosenbaum PR. Optimal matching of an optimally chosen subset in observational studies. *J Comput Graph Stat*. 2012;21(1):57-71.
- Athey S. The impact of machine learning on economics. In: Agrawal AK, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda*. Chicago, IL: University of Chicago Press; 2018:507-532.
- Mullainathan S, Spiess J. Machine learning: an applied econometric approach. *J Econ Perspect*. 2017;31(2):87-106.
- Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care*. 2010;48(6):S106-S113.
- Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216-1219.
- Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis*. 2018;66(1):149-153.
- Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17.
- Efron B, Hastie T. *Computer Age Statistical Inference*. Cambridge University Press; 2016.
- Rose S. A machine learning framework for plan payment risk adjustment. *Health Serv Res*. 2016;51(6):2358-2374.
- Rokach L. Ensemble-based classifiers. *Artif Intell Rev*. 2010;33(1-2):1-39.
- Polley EC, Rose S, van der Laan MJ. Super learning. In: Polley EC, van der Laan MJ, eds. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer; 2011.

27. Rose S. Mortality risk score prediction in an elderly population using machine learning. *Am J Epidemiol*. 2013;177(5):443-452.
28. Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. Prediction policy problems. *Am Econ Rev*. 2015;105(5):491-495.
29. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
30. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat*. 2006;2(1):1-40.
31. van der Laan M, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer; 2011.
32. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20(1):217-240.
33. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Statist*. 2019;47(2):1148-1178.
34. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113(523):1228-1242.
35. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. Double/debiased/Neyman machine learning of treatment effects. *Am Econ Rev*. 2017;107(5):261-265.
36. Hahn PR, Murray JS, Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (Mimeo). 2018. <https://math.la.asu.edu/~prhahn/BCF.pdf>. Accessed November 1, 2018.
37. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185(1):65-73.
38. Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Stat Med*. 2018;37(16):2530-2546.
39. Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat Methodol*. 2005;2(2):131-154.
40. Rose S, van der Laan MJ. Simple optimal weighting of cases and controls in case-control studies. *Int J Biostat*. 2008;4(1):19.
41. van der Laan MJ. Estimation based on case-control designs with known prevalence probability. *Int J Biostat*. 2008;4(1):17.
42. Fielding A, O'Muircheartaigh CA. Binary segmentation in survey analysis with particular reference to AID. *J Roy Stat Soc Ser D Stat*. 1977;26(1):17-28. <https://doi.org/10.2307/2988216>.
43. Messenger R, Mandell L. A modal search technique for predictive nominal scale multivariate analysis. *J Am Stat Assoc*. 1972;67(340):768-772.
44. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat*. 2010;4(1):266-298.
45. Hahn PR, Carvalho CM, Puelz D, He J. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal*. 2018;13(1):163-182.
46. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
47. Frisch R, Waugh FV. Partial time regressions as compared with individual trends. *Econometrica*. 1933;1(4):387-401.
48. Lovell MC. Seasonal adjustment of economic time series and multiple regression analysis. *J Am Stat Assoc*. 1963;58(304):993-1010.
49. Angrist JD, Krueger AB. Split-sample instrumental variables estimates of the return to schooling. *J Bus Econ Stat*. 1995;13(2):225-235.
50. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*, 2nd edn. New York, NY: Springer; 2009.
51. Belloni A, Chen D, Chernozhukov V, Hansen C. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*. 2012;80(6):2369-2429.
52. Hartford J, Lewis G, Leyton-Brown K, Taddy M. Deep IV: a flexible approach for counterfactual prediction. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. JMLR.org; 2017:1414-1423. <http://dl.acm.org/citation.cfm?xmid=3305381.3305527>. Accessed August 1, 2019.
53. Gilchrist DS, Sands EG. Something to talk about: social spillovers in movie consumption. *J Polit Econ*. 2016;124(5):1339-1382.
54. Windmeijer F, Farbmacher H, Davies N, Smith GD. On the use of the lasso for instrumental variables estimation with some invalid instruments. *J Am Stat Assoc*. 2018;14:1-12. <https://doi.org/10.1080/01621459.2018.1498346>
55. Lu C, Nie X, Wager S. Robust nonparametric difference-in-differences estimation. arXiv:190511622 [stat]. 2019. <http://arxiv.org/abs/1905.11622>. Accessed August 1, 2019.
56. Abraham S, Sun L. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Working Paper. 2018.
57. Zheng W, van der Laan MJ. Asymptotic theory for cross-validated targeted maximum likelihood estimation (UC Berkeley Division of Biostatistics Working Paper 273). 2010.
58. Yang Y, Cheng G, Dunson D. Semiparametric Bernstein-von Mises theorem: second order studies (ArXiv:1503.04493). 2015. <https://arxiv.org/abs/1503.04493>. Accessed March 28, 2019.
59. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econometr J*. 2018;21(1):C1-C68.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: McConnell KJ, Lindner S. Estimating treatment effects with machine learning. *Health Serv Res*. 2019;54:1273-1282. <https://doi.org/10.1111/1475-6773.13212>