Optimal hyperparameter tuning of random forests for estimating causal treatment effects

Article *in* Songklanakarin Journal of Science and Technology · September 2021

DOI: 10.14456/sjst-psu.2021.132

3 authors:

Lateef Amusa
University of Johannesburg
34 PUBLICATIONS **69** CITATIONS
SEE PROFILE

Delia North
University of KwaZulu-Natal
84 PUBLICATIONS **423** CITATIONS
SEE PROFILE

Temesgen T Zewotir
University of KwaZulu-Natal
188 PUBLICATIONS **1,574** CITATIONS
SEE PROFILE

Songklanakarin J. Sci. Technol.
43 (4), 1004-1009, Jul. - Aug. 2021

http://www.sjst.psu.ac.th

Original Article

Optimal hyperparameter tuning of random forests for estimating causal treatment effects

Lateef Amusa*, Delia North, and Temesgen Zewotir

Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Received: 4 June 2020; Revised: 15 July 2020; Accepted: 23 July 2020

Abstract

Recent studies have expanded the focus of machine learning methods like random forests beyond prediction. They have found utility in the area of causal inference by using it to estimate propensity scores. It has also been established in the literature that tuning the hyperparameter values of random forests can improve the estimates of causal treatment effects. We thus address the issue of getting the best out of random forest models by proposing to tune the random forest hyperparameters while maximizing covariate balance. We consider variants of tuning based on a model fit criterion and compare with tuning to chase covariate balance. In a simulation study and empirical application in two case studies, we studied the performance of different tuning implementations, relative to the random forest with default hyperparameters. We find that tuning to chase balance rather than model fit when estimating propensity scores induced better balance in the covariates and produced more accurate treatment effect estimates.

Keywords: random forest, simulation, observational studies, propensity scores, treatment effect, causal inference

1. Introduction

Randomized control trials (RCTs) are typically the best research design to learn if a treatment is effective. However, a well-designed randomized trial may be neither ethical nor affordable to conduct. In many applied studies, there is an increasing interest in the utilization of observational data for estimating causal treatment effects. Unlike RCTs, the baseline characteristics of the treatment groups often differ systematically in observational studies, thereby introducing selection bias or confounding.

Propensity score (PS) methods have by far been the most popular approach to minimizing confounding when estimating causal treatment effects (Austin, 2014; Dehejia & Wahba, 2002; Guo, Barth, & Gibbons, 2006; Guo & Fraser, 2010; Hirshberg & Zubizarreta, 2017). The popularity of propensity scores gave rise to several methodological approaches to its estimation. These methods include, but are not limited to, logistic regression, machine learning methods (Breiman, 2001; Hill, 2011; McCaffrey, Ridgeway, & Morral, 2004; Pirracchio, Petersen, & van der Laan, 2015), entropy balancing (Hainmueller, 2012), and covariate balancing propensity scores (Imai & Ratkovic, 2014).

In recent times, machine learning techniques, which can be viewed as best suited for prediction problems, have also been expanded into the area of statistical inference; i.e. estimating treatment effects with corresponding precision. Because machine learning techniques are data-adaptive and do not require any prior assumptions about the correct functional form of the model, they do not rely on the correct specification of the PS model. Further, some studies (Austin, 2012; Lee, Lessler, & Stuart, 2010) showed that ensemble-based methods, which include random forests (Breiman, 2001) and generalized boosted models (McCaffrey *et al.*, 2004) outperformed the traditional logistic regression for estimating propensity scores.

Machine learning algorithms involve several hyperparameters that control their model complexities and performance. Hyperparameters are preset parameter values of a machine learning algorithm. A systematic selection of these

*Corresponding author
Email address: amusasuxes@gmail.com

hyperparameters, also referred to as hyperparameter tuning, is usually done to improve model fit. Though hyperparameter tuning has been extensively studied for prediction problems, clear guidance is missing for causal inference problems. In this study, we mainly focus on random forests. Random forests, like generalized boosted models, have been relatively more utilized in estimating propensity scores, especially in the context of PS weighting.

The standard method for tuning random forest models, like other machine learning techniques, is to select the hyperparameters which yield the best model fit. Specifically, for random forests, the best model fit usually refers to the smallest out-of-bag prediction error estimated using cross-validation or a holdout sample. Another approach we are proposing is to set the tuning parameter so that the resulting random forests and associated weights minimize covariate imbalance between treatment groups.

It is pertinent to provide clarity on how model tuning impacts covariate balance, as well as accuracy and precision of treatment effect estimates. Thus, this paper aims to determine the optimal tuning of the random forest model in terms of estimation of causal treatment effects. We addressed the objective with a small-scale simulation study, and previously illustrations with two empirical case studies.

2. Materials and Methods

2.1 Overview of random forests

Random forests are a type of ensemble-based methods, which builds on the classification and regression trees (CART) algorithm for growing unit trees. Further details on CART can be found elsewhere (James, Witten, Hastie, & Tibshirani, 2013; Lemon, Roy, Clark, Friedmann, & Rakowski, 2003). Random forests aggregate many trees into a robust ensemble by taking repeated bootstrap samples from the study sample. It then grows a CART-like tree in each of these bootstrap samples. However, the respective individual trees are restricted to consider only a random subset *m* of the *p* predictors at each split point. For each study subject or units, the estimated outcome is then obtained from averaging the predictions from the grown trees (if the goal is regression) or from a majority vote (if classification is the goal).

2.2 Estimation of propensity score weights with random forests

Let us assume a binary indicator variable T of treatment, and $\boldsymbol{X}$ is a vector of observed covariates. The propensity score, defined by $\pi(\boldsymbol{X}) = P(T = 1 \mid \boldsymbol{X})$, $0 < \pi(\boldsymbol{X}) < 1$, is the probability of a subject or unit receiving the treatment of interest, given the observed baseline covariates. The propensity scores $\pi(\boldsymbol{X})$ are estimated as prediction probabilities from the random forest procedure described in the preceding section. Additionally, for clarity, a CART model of the covariates $\boldsymbol{X}$ on treatment assignment T is estimated using the data from each random sample selected with replacement. The units or subjects' propensity scores are estimated from each CART model. These propensity scores are then averaged across all the individual decision tree models to obtain the random forest propensity score for each participant. Since the treatment variable is binary, our random forest model was based on a classification tree; hence, the propensity scores are estimated from the fitted classification trees as classification probabilities of the most occurring class. Under the assumption of selection on observables, we can then use $\pi(\boldsymbol{X})$ to estimate the causal estimand of interest.

Our causal estimands of interest in this study are the average treatment effect on the population (ATE) and the average treatment effect among the treated (ATT). Given that we aim to utilize propensity scores in weighting, the ATE weights, also known as the inverse probability of treatment weighting, was given by $\frac{1}{\pi(\boldsymbol{X})}$ for the treated group, while $\frac{1}{1-\pi(\boldsymbol{X})}$ was assigned for the control group. ATT weights, also known as weighting by the odds, were assigned a value of 1 for the treated group, while the weights of the control group were estimated as $\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}$.

2.3 Tuning the random forest model for the estimation of treatment effects

Tuning is the process of searching the optimal hyperparameters of a learning algorithm for a considered dataset. Though several hyperparameters control the randomness of random forests, 3 basic parameters stand out: *mtry*, node size, and sample size. Out of these 3 parameters, *mtry* has the most substantial impact (Probst, Wright, & Boulesteix, 2019; van Rijn & Hutter, 2018). *mtry* denotes the number of a randomly drawn subset of variables considered at each split in the tree. The node size is the minimum number of observations in a terminal node. The sample size is the number of randomly drawn observations for training each tree.

The number of trees is not tunable in the classical sense; it should be set as high as possible. Other parameters are left in their default values of the randomForest package (Liaw & Wiener, 2002). For example, the splitting rule consists of selecting, out of the *mtry* candidate variables, the split that minimizes the Gini impurity (Berk, 2005).

In tuning for model fit, we applied an automatic tuning procedure that iteratively assessed the cross-validated performance of the random forest over a range of plausible *mtry* values. We then chose the *mtry* value that yielded the minimum cross-validated out-of-bag prediction error. This was done using the tuneRF function of the randomForest package (Liaw & Wiener, 2002) in R (Team, 2016).

As stated in the introduction section, we can alternatively tune for maximizing covariate balance. It involves specifying a metric for assessing covariate balance. Here, we utilized the absolute standardized mean difference (ASMD) between the two treatment groups for each covariate. Some authors suggested that ASMD values above 0.1 may be indicative of covariate imbalance (Mamdani *et al.*, 2005; Normand *et al.*, 2001). Depending on whether the ATE or ATT is of interest, the ASMD is standardized by the standard deviation of the pooled sample, and the standard deviation of the treated group only. In tuning for optimal covariate balance, we select the *mtry* value that minimized the maximum ASMD of the covariates. We achieved this with the randomForest package, together with our user-defined function in R.

2.4 Simulation study

We conducted Monte Carlo simulation experiments to examine the performance of the different tuned random forest models, relative to random forests using default hyperparameters. We replicated a data generating process (DGP) with settings similar to previous studies that provide a flexible simulation structure and mimics practical problems (Abdia, Kulasekera, Datta, Boakye, & Kong, 2017; Leacy & Stuart, 2014). Our DGP assumed there were 10 multivariate normal distributed baseline covariates $\boldsymbol{X} = (X_1, \ldots, X_{10})$, with zero means, unit variance, and varying degrees of correlation (0.2 and 0.9) between pairs of covariates. It involves complex and non-linear relationships between the treatment indicator T and $\boldsymbol{X}$, as well as outcome Y and $\boldsymbol{X}$. The assumed true propensity score function was generated as

$$\text{Logit } \pi(\boldsymbol{X}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_{11} X_1^2 + \alpha_{22} X_2^2 + \alpha_{23} X_2 X_3 + \alpha_{345} X_3 X_4 X_5, \quad (1)$$

where $(\alpha_1, \alpha_2, \alpha_3, \alpha_{11}, \alpha_{22}, \alpha_{23}, \alpha_{345})$ = (log (1.25), log (1.5), log (1.75), log (1.25), log (1.5), log (1.75), log (2)), which introduced varying small to large effect sizes, and α_0 was chosen to ensure that approximately 33% of the population received the treatment. Outcome model was of the form:

$$Y = \beta_0 + \beta_1 X_1^2 T + \beta_2 X_4 T + \beta_3 X_1 X_4 (1- T) + \beta_4 X_5 (1- T) + \varepsilon_i, \varepsilon_i \sim N\left(0, \frac{\text{Var } E(Y/X)}{50}\right),$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ = (0, 2, 3, 2, -4).

We simulated 100 datasets of size 1000 and estimated the propensity scores from the different methods linearly using all the 10 covariates, even though only a subset impacted the true propensity scores.

We estimated both the ATE (corresponding true ATE = 1.99) and ATT (corresponding true ATE = 2.536). For each method, we calculated the bias (% deviation from the true estimate), root mean squared error (RMSE), standard deviation, model-based standard error, and the 95% nominal coverage rates (the proportion of times the estimated 95% confidence interval includes the true treatment effect).

3. Results

We present the results for the simulation experiments according to each of the performance metrics explained in the earlier section. We present results under each of the ATT and ATE. Figure 1 shows show box plots for the maximum ASMD from the ten covariates used in the simulation DGP. Our proposed method of tuning to chase balance performed best in terms of balance, with maximum ASMDs generally lower (range: 0.005 – 0.115; median: 0.043), regardless of whether interest lies in estimating ATE or ATT. However, model fit tuning was much better than no tuning. Better covariate balance was generally achieved when ATE was of interest than when ATT was estimated.

When estimating ATE, Table 1 shows that tuning by chasing balance outperformed the alternatives in terms of bias, standard deviation, and RMSE of the estimated treatment effects. The biggest accuracy gain in the two tuning strategies, relative to the default random forest model was in the bias of ATE. Specifically, tuning by chasing balance and tuning for model fit reduced the bias by 64% and 45%, respectively. Though there were no substantial differences in the random forests models in terms of standard errors and 95% coverage rates, tuning for model fit produced the smallest standard error.

When estimating ATT, substantial differences in the tuning strategies were only observed in the bias. Model fit tuning appeared to have competed favourably in this setting; however, model fit tuning did not substantially outperform tuning to chase balance.

4. Empirical examples

We present and discuss results from two distinct examples estimating ATE and ATT, which illustrate the utility

Table 1. Treatment effect estimates from the simulation study

		Default value	Tuned value (model fit)	Tuned value (chasing balance)
ATE	% bias	9.54	3.43	5.23
	RMSE	0.172	0.158	0.171
	Std.Dev	0.143	0.155	0.163
	Mean SE	0.326	0.326	0.32
	95% CI coverage	100	100	100
ATT	% bias	5.58	7.36	6.82
	RMSE	0.236	0.269	0.263
	Std.Dev	0.23	0.259	0.255
	Mean SE	0.394	0.394	0.38
	95% CI coverage	100	100	99

Note: The results are averaged over the simulation runs

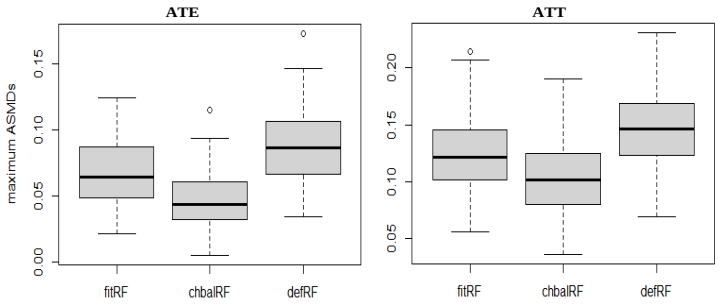

Figure 1. Maximum ASMD of the covariates in the simulation study. The results are averaged over the simulation runs.

of tuning random forests with different approaches to achieve accurate and precise estimation of treatment effects. Due to computational simplicity, we tuned only the *mtry* parameter in the simulation study. However, here, we extended our tuning for model fit to a sequential model-based optimization (SMBO) procedure, simultaneously tuning the 3 parameters *mtry*, sample size and node size. We chose the area under the ROC curve (AUC) as our performance metric to be optimized. This was achieved using the more comprehensive R tuneRanger package (Probst *et al.*, 2019). In summary, we compared three model fit tuning strategies, namely, tuning *mtry* only for model fit (tuneRF1), SMBO tuning for model fit (tuneRF2), and tuning to chase balance (balRF). We used the default hyperparameters of the random forest model (defRF) to benchmark our results.

Like the simulation study, we estimated propensity scores from the different methods using all available covariates. For each tuning strategy, 1000 trees were used to train the random forest. Performance evaluation was based on covariate balance and outcome estimation.

4.1 Case study for ATE

We used the different tuning strategies to estimate ATE by reanalyzing the Lindner dataset from the R package *twang (*Ridgeway *et al.*, 2020*)*. The dataset comprises information on 996 patients who received an initial Percutaneous Coronary Intervention (PCI) received at the health facility at that time. The treated group are patients who received the PCI with additional treatment, abciximab - an expensive, high-molecular-weight IIb/IIIa cascade blocker, while the control group are those who received the PCI alone. Covariates include height, number of vessels involved in initial PCI (ves1proc), an indicator for recent acute myocardial infarction (acutemi), left ventricle ejection fraction (ejecfrac), an indicator for coronary stent insertion (stent), a diabetic indicator (diabetic), and gender (female). One of the outcome variables was the treatment cost for the first 6 months. The Lindner study aimed to determine the cost-effectiveness of abciximab.

Optimal hyperparameter values for the different random forests are as follows: *tuneRF1* (*mtry* = 2), *tuneRF2* (*mtry* = 5, node size = 17, sample size = 0.514), *balRF* (*mtry* = 2). For tuneRF1 and balRF, their corresponding hyperparameters coincided. Figure 2 shows that the three tuning strategies performed remarkably well in reducing covariate imbalances, with ASMD values well below the 0.1 threshold. However, *balRF* and *tuneRF2* both performed best (max ASMDs = 0.037). Our estimates for the cost difference of the first 6 months after treatment was roughly between 623 and 763 dollars, although *tuneRF2* had a higher standard error (Table 2). Hence, regardless of the method used to tune the random forest models, the abciximab treatment of the first six months appears to increase cost, although confidence intervals include zero.

4.2 Case study for ATT

To illustrate the estimation of ATT, we used data from a merger of the 185 treated group participants from the experimental National Supported Work Demonstration (NSW) program (LaLonde, 1986) and the 15992 Current Population Survey (CPS) control group participants. The dataset, which can be found in https://users.nber.org/~rdehejia/data/.nswdata2.html, included the following covariates: real earnings in 1974 (*re74*) and 1975 (*re75*), age (*age*), number of years of education (*edu*), indicator variables for unemployment in 1974 (*u74*) and 1975 (*u75*), marital status (*married*), hispanic race (*hisp*), and black race (*black*), no high school diploma/degree (*nodeg*). The NSW program aimed to determine if the postintervention earnings increased. Optimal hyperparameter values for the different random forests are as follows: *tuneRF1* (*mtry* = 2), *tuneRF2* (*mtry* = 3, node size = 6, sample size = 0.406), *balRF* (*mtry* = 5). Figure 3 shows that only *balRF* successfully induced balance in all

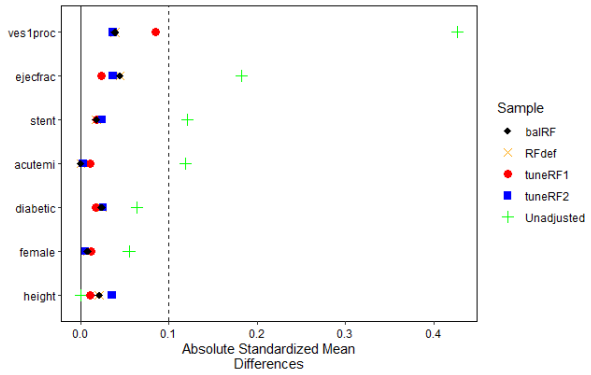

Figure 2. Covariate balance assessment in the Lindner case study

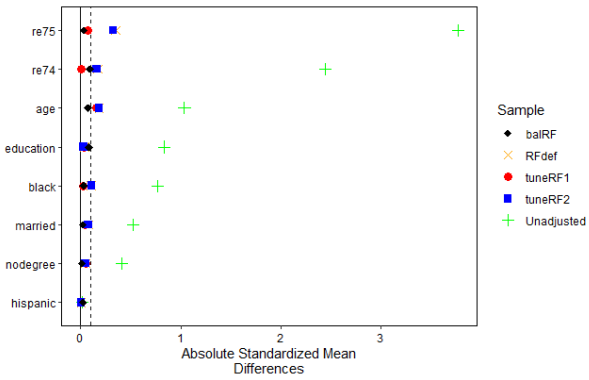

Figure 3. Covariate balance assessment in the Lalonde case study

Table 2. Treatment effect estimates for the Lindner case study

	Rfdef	tuneRF1	tuneRF2	balRF
Estimated treatment effect	688	623	763	688
SE	986	1024	1032	986
Maximum ASMD	0.045	0.037	0.086	0.037
95% CI	[-1243, 2620]	[-1383, 2629]	[-1260, 2786]	[-1383, 2629]

Table 3. Treatment effect estimates for the Lalonde case study

	Rfdef	tuneRF1	tuneRF2	balRF
Estimated treatment effect	552	612	1334	1812
SE	635	635	674	830
Maximum ASMD	0.348	0.325	0.152	0.094
95% CI	[-693, 1796]	[-633, 1857]	[13, 2655]	[185, 3439]

the covariates, with ASMD values well below 0.1)max ASMD = 0.094(. Our estimates for the increased difference in earnings for the year 1978 was in the range of 635 and 830. Like the simulation results, *balRF* had the highest standard error)Table 3(. Only *balRF* and *tuneRF2* suggest that among the participants assigned to the job training group, the job training intervention effect on earnings was statistically significant.

5. Discussion

In this study, we examine an important issue in the implementation of random forests to estimate propensity score weights: optimal tuning of the hyperparameters of a random forest model. Do we tune to optimize model fit or tune to optimize covariate balance? Results from our simulation and reanalysis of data from two case studies suggest that tuning random forest models to obtain the best model fit does not necessarily result in the best balance in the treatment groups and bias reduction in the treatment effect estimates. These findings support the findings of previous studies (Griffin, McCaffrey, Almirall, Burgette, & Setodji, 2017; Westreich, Cole, Funk, Brookhart, & Stürmer, 2011).

In terms of accuracy of treatment effect estimates, our results favoured tuning to chase balance when estimating ATE. At the same time, our findings are less clear for ATT estimation. However, in terms of precision, the results for inferences are less clear. Whether the interest lies in estimating ATE or ATT, no method of tuning random forests provides smaller standard errors and better coverage than the others. Thus, even though the results on inferences do not favour any tuning strategy, tuning to chase covariate balance may still be preferable because it produces more accurate treatment effect estimates.

This study has its caveats. First, our simulations were relatively simple and limited. There is a need to extend our simulation study with different but challenging data-generating processes and scenarios. In particular, the effect of increasing confounders, sample size, and noise variables could change the direction of this study. Secondly, when tuning for model fit in the empirical examples, other performance measures could have been used. For example, in the sequential model-based optimization (SMBO) tuning, Brier score or logarithmic loss could potentially replace AUC.

As far as we know, this is the very first study to investigate the optimal tuning of random forests in the area of causal inference. Future studies may consider the use of the super learning methodology, which runs a weighted combination of several machine learning algorithms to estimate propensity scores (Pirracchio *et al.*, 2015; van der Laan, Polley, & Hubbard, 2007). It might further be of interest to study the gains in improving the performance of this method by tuning it to select the optimal combination of machine-learners that yields the best balance.

6. Conclusions

In summary, we found that tuning random forests to estimate causal treatment effects warrants consideration in applied analyses. In particular, the proposed method of tuning to chase balance resulted in estimates of average treatment effects with low bias.

References

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal, 59*(5), 967-985.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate behavioral research, 47*(1), 115-135.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33*(6), 1057-1069.

Berk, R. (2005). An introduction to ensemble methods for data analysis. Paper 2005032701, Department of Statistics Papers, 2005. Retreived from: http://repositories.cdlib.org/uclastat/papers/2005032701.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151-161.

Griffin, B. A., McCaffrey, D. F., Almirall, D., Burgette, L. F., & Setodji, C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of Causal Inference, 5*(2).

Guo, S., Barth, R., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review, 28*, 357–383.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: SAGE Publications.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis, 20*(1), 25-46.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20*(1), 217-240.

Hirshberg, D. A., & Zubizarreta, J. R. (2017). On Two Approaches to weighting in causal inference. *Epidemiology, 28*(6), 812-816.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(1), 243-263.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (Volume 112)*. Berlin, Germany: Springer.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604-620.

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

Lee, Lessler, & Stuart. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346.

Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine, 26*(3), 172-181.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2*(3), 18-22.

Mamdani, M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., . . . Anderson, G. M. (2005). Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *Bmj, 330*(7497), 960-962.

McCaffrey, Ridgeway, & Morral. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*, 403-425.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*(4), 387-398.

Pirracchio, R., Petersen, M. L., & van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology, 181*(2), 108-119.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), e1301.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., Burgette, L. & Cefalu, M. (2020). Twang: Toolkit for weighting and analysis of nonequivalent groups. R package version 1.6 Retreived from https:// CRAN.R-project.org/package=twang

Team, R. C. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

van der Laan, M., Polley, E., & Hubbard, A. (2007). Super learner. *Statistical Applications of Genetics and Molecular Biology, 6*, Article 25.

van Rijn, J. N., & Hutter, F. (2018). Hyperparameter importance across datasets. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., & Stürmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety, 20*(3), 317-320.