Published in final edited form as:

Epidemiology. 2017 November ; 28(6): 802–811. doi:10.1097/EDE.0000000000000734.

The right tool for the job: choosing between covariate balancing and generalized boosted model propensity scores

Claude M. Setodji, PhD¹, Daniel F. McCaffrey, PhD², Lane F. Burgette, PhD¹, Daniel Almirall, PhD³, and Beth Ann Griffin, PhD¹

¹RAND Corporation, Pittsburgh, PA 15213 USA

²Educational Testing Service, Princeton, NJ 08541, USA

³Institute for Social Research, University of Michigan, Ann Arbor, MI 48104, U.S.A

Abstract

Estimating the causal effect of an exposure (versus some control) on an outcome using observational data often requires addressing the fact that exposed and control groups differ on pre-exposure characteristics that may be related to the outcome (confounders). Propensity score methods have long been used as a tool for adjusting for observed confounders in order to produce more valid causal effect estimates under the strong ignorability assumption. In this article, we compare two promising propensity score estimation methods (for time invariant binary exposures) when assessing the average treatment effect on the treated: the generalized boosted models and covariate-balancing propensity scores, with the main objective to provide analysts with some rules-of-thumb when choosing between these two methods. We compare the methods across different dimensions including the presence of extraneous variables, the complexity of the relationship between exposure or outcome and covariates, and the residual variance in outcome and exposure. We found that when non-complex relationships exist between outcome or exposure and covariates, the covariate-balancing method outperformed the boosted method, but under complex relationships, the boosted method performed better. We lay out criteria for when one method should be expected to outperform the other with no blanket statement on whether one method is always better than the other.

Keywords

Causal inference; observational study; propensity score; variable selection; diagnostic tools

Introduction

Epidemiologic studies often focus on estimating the impact of an exposure of interest on a health outcome, such as the impact of radon exposure on lung cancer¹ or of second hand

Address correspondence to: Claude M. Setodji, RAND Corporation, 4570, 5th Avenue, Suite 600, Pittsburgh, PA, 15213-2665, setodji@rand.org, Phone : 412-683-2300, Fax : 412-683-2800.

Conflict of Interest: All the authors have no conflicts of interest to disclose.

Authors' contribution: All the listed authors were part of the team that conceptualized and designed the study. They all drafted different parts of the initial manuscript, reviewed or revised it and approved the final manuscript as submitted.

smoking on heart diseases and mortality². Because the processes that determine individuals' exposures may also directly impact their outcomes of interest, great care is often taken in the analysis of observational data to ensure that exposure effect estimates reflect a true causal association rather than merely the effects of confounding. Propensity score methods are widely used in epidemiology and other fields to improve causal effect estimates by balancing exposed and unexposed groups based on a rich set of pre-exposure characteristics. For time-invariant dichotomous exposures (or treatments, in the more common terminology), the critical step in propensity score analyses is the estimation of the propensity score itself, which is defined as the probability that each individual is exposed or receives the treatment versus not, as a function of observed pre-treatment covariates^{3;4}. In recent years, methodological advances have produced propensity score estimation methods that substantially improve upon the standard logistic regression models that underpinned the original work in the field. Comparative studies e.g.,^{5;6} have found that two promising methods are generalized boosted models (GBM)⁷ and covariate balancing propensity scores (CBPS)⁸. However, the literature gives practitioners little information on which empirical situations call for which of these two methods.

Through extensive simulations — involving some 4 million simulated datasets — this paper seeks to give analysts a measure of clarity when choosing between the two methods in the case of time-invariant exposure or treatment. As opposed to claiming that one method is better than the other (which we do not believe is true), this paper lays out both qualitative and quantitative criteria for when one method may be expected to outperform the other. In the end, we believe that both methods should be in the toolbox of epidemiologists. Our goal is to offer context-dependent advice on which tool to reach for first and also provide performance results. This study was approved by the Human Subjects Protection Committee at the RAND Corporation.

Background

Inverse probability weighting for time-invariant binary treatment

Assuming a time-invariant binary treatment variable of interest T taking value 1 if a study participant is in the treatment and 0 otherwise, and assuming a set of confounders $\mathbf{X} = (X_1, X_2, \dots, X_k)$, the propensity score $p(\mathbf{x})$ is defined as the probability of being assigned to the treatment conditional on $\mathbf{X}$ or formally $p(\mathbf{x}) = \text{Prob}(T = 1 | \mathbf{X} = \mathbf{x})$. Strong ignorability of treatment assignment⁴ states that (1) all confounders are captured in $\mathbf{X}$ and (2) there is a positive probability of receiving the treatment for all values of $\mathbf{X}$ ($0 < p(\mathbf{x}) < 1$ for all $\mathbf{x}$). Under this assumption, the propensity score is all that is needed to control for pretreatment differences between the treatment and the control groups, which may be operationalized through matching, weighting or subclassification⁹.

This study focuses on weighting comparisons since both boosted model and covariate balancing methods were developed with optimizing weighted balance in mind. With weighting, once an estimate $\hat{p}(\mathbf{x})$ of $p(\mathbf{x})$ is obtained, a function of the propensity score is used as a weight in the estimation of the difference in the outcome between the treatment and control groups, a difference that will be a consistent estimator of the treatment effect under the strong ignorability assumption. The most common estimands of the treatment

effect are the average treatment effect on the population of interest and the average treatment effect on the treated. For average treatment effect, the weight equals $\frac{T}{\hat{p}(x)} + \frac{1-T}{1-\hat{p}(x)}$, while for average treatment on the treated, it equals $T + \frac{(1-T)\hat{p}(x)}{1-\hat{p}(x)}$ ^{10;11;12}. In this study, we will conduct comparisons between weighted estimators using boosted models versus covariate balancing method for estimating the average treatment on the treated.

Propensity score estimation using covariate balancing propensity scores

To estimate the propensity score, researchers typically posit a parametric model that is linear in the unknown parameters and is fit via maximum likelihood. The logistic model is most popular, where

$$p(x) = \text{Prob}(T=1 | \mathbf{X}=x) = \frac{\exp(\boldsymbol{\beta}^\top \mathbf{X})}{1 + \exp(\boldsymbol{\beta}^\top \mathbf{X})}.$$

One of the limitations of parametric modeling is that the model may be misspecified. The covariate balancing method developed by⁸, which models exposure while optimizing the covariate balance, confers robustness to mild model misspecification with regard to balancing confounders compared to direct maximum likelihood estimation. Further, even if the model is correctly specified, the covariate balancing method can improve the covariate balance in observed data sets and potentially improve the accuracy of estimated treatment effects over traditional logistic regression.

In addition to maximizing the model likelihood, the covariate balancing method incorporates a balance condition for the weighted means of the covariates in the parameter estimation procedure. Covariate balancing uses a generalized method of moments or an empirical likelihood estimation framework^{13;14} to find estimates that come closest to optimizing the likelihood function while meeting the balance condition simultaneously.

Propensity score estimation using generalized boosted models

Generalized boosted model is a nonparametric, piecewise constant model for predicting the treatment T^7 . This method builds the propensity score model iteratively, starting from a globally constant model, and gradually increasing model complexity. At each step, a simple regression tree¹⁵ is added to the current model gradually creating an increasingly complex piecewise constant function. In propensity score applications, generalized boosted model's complexity is tuned by optimizing covariate balance between the inverse-probability-weighted treatment and control samples^{16;17}. The boosted model's tree-based approach avoids making any assumptions of linearity in unknown parameters, and can automatically accommodate interactive effects in the propensity score model. In addition to flexibility, by using a piecewise constant approximation to the propensity score, the boosted models method estimates 'flatten out' over areas of the covariate space where few observations are available, often resulting in more stable propensity score estimates.

Study design

To assess the performance of the standard implementations of covariate balancing propensity scores and generalized boosted models for estimating propensity scores, we conducted a simulation study to test conjectures about the impact of five factors on the relative accuracy of treatment effects estimated using weights derived from propensity scores estimated by the boosted model or the covariate balancing method (see Table 1). In Table 1, extraneous variables refer to variables used in the modeling but unrelated to either the treatment (“outcome-only predictors”) or the outcome (“instrumental variables”) or both (“distractors”). When the covariates are weakly related to treatment we expect both methods to perform well because the treatment and control groups are highly similar even without weighting. However, given the non-parametric aspect of the boosted models method, it may possibly suffer inefficiencies in these circumstances while trying to fit possibly non-linear relationships that are better approximated via linear ones.⁵

Data generation

Simulated predictors—We used a modification of the simulation structure described by Setoguchi et al. (2008) and replicated by others (Wyss et al., 2014; Lee et al., 2009). For each simulation, we generated 15 predictors as a mixture of continuous and binary variables (see more details in eAppendix Table 1). The continuous predictors were standard normal and the binary variables were dichotomized (at cut point 0) versions of standard normal random variables.

Simulation of treatment—We generated treatment assignments using one of seven versions for the propensity score model. All seven versions were of the form $\text{Prob}(T=1) = 1/\{1 + \exp(-\text{version} - \tau\xi)\}$, where *version* was a function of the confounders that established the complexity of the association between them and treatment assignment. The versions vary from simple linear combinations of covariates up to very complex relationships with non-linearity and interaction between covariates (see Table 2). The parameters $\alpha_1, \alpha_2, \dots, \alpha_{24}$ have values between -0.8 and 0.8 . The variable $\xi \sim N(0, 1)$ controls the strength of the covariates for predicting treatment through τ which takes one of six value $\tau = 0, 0.25, 0.5, 1, 1.5, 2$. For each version, smaller values of τ correspond to a stronger relationship between treatment assignment and the covariates.

Simulation of outcome—We generated outcomes $Y = m(\mathbf{X}) + \sigma\zeta$, using five different models for $m(\mathbf{X}) = E[Y | \mathbf{X}]$ where $\mathbf{X} = (X_1, X_2, X_3, X_4, X_8, X_9, X_{10})'$ ranging from linear additive to nonlinear and non-additive (see Table 3). The parameters $\delta_1, \delta_2, \dots, \delta_7$ had values between -0.73 and 0.71 and the intercept $\delta_0 = -3.85$. The treatment effect θ was assumed additive and equal to -0.4 . For each outcome specification, $\zeta \sim N(0, 1)$ and 20 different σ were considered ranging from 0 to 5 (0.25 point increments) to control the amount of error variance.

Overall, the simulation scenarios were composed of seven versions crossed with six different levels of random variability to specify the propensity score model and five outcome models with 20 different error variances for a total of 4, 200 scenarios. For each scenario we

generated 1, 000 simulated data sets and used the methods being compared to estimate propensity scores and treatment effects.

Estimation

The propensity estimation strategies were designed to assess the impact of the set of variables included in the propensity score estimation model. See a detail description of the strategies in eAppendix Table 1. In practical situations, researchers make a decision on what variable to include in the estimation procedures, and these strategies mimic some of those decisions. The options of variables included considered were:

1. Only confounders directly related to both the outcome and treatment
2. Variables only directly associated with the treatment, including the instrument X_7
3. Only variables associated directly to the outcome
4. All variables directly or indirectly related to both outcome and treatment
5. All variables related to the outcome or to the treatment
6. All available variables including distractors

For each strategy, we used both the boosted model and the covariate balancing methods to estimate the propensity score models, which were then used to generate the average treatment on the treated inverse probability weights.

Comparison of generalized boosted models and covariate balancing propensity scores methods

Performance and comparison measures—To evaluate the performance of both methods on the different propensity score versions being considered, we compared the measures on how well they balanced the covariate distributions between the treated and non-treated groups and on the bias and accuracy of the resulting treatment effect estimates defined as follows:

The Average Standardized Absolute Mean Difference (referred to as standardized mean difference henceforth) measures differences in group means of the covariates. To calculate the standardized mean difference, for each covariate, we first computed the absolute value of the difference between the average treatment on the treated weighted treatment and control group means, then standardized this difference by the standard deviation of the covariate in the treatment group and finally averaged these values across all covariates.

The Kolmogorov-Smirnov Statistic is a nonparametric measure of the difference in the entire univariate distribution of confounders between the treatment and the control group and is defined as the largest absolute difference in the weighted empirical cumulative distribution function of the groups. Because it is sensitive to any difference in the entire distribution of a covariate, rather than just the mean as in the standardized mean difference, it can be important for detecting imbalance when non-linear transformations of confounders are in the the confounding scheme. For the

Kolmogorov-Smirnov statistic and the standardized mean difference, balance is calculated for each simulated dataset and averaged over datasets within a scenario.

The variance of the weights has the potential to increase the variability of estimated treatment effects and can indicate large influence for a small number of observations. For each simulated dataset we estimated the sample standard deviation of the weights for the control group and report the average over the datasets in scenario.

Bias is the difference between the expected value of the estimated treatment effect and the true effect set at -0.4 . We report the absolute value of the average over the simulated datasets in each scenario.

Standard error is the standard deviation of the estimated treatment effects across the simulated datasets in each scenario.

Root mean squared error (RMSE) is the square root of the expected value of the square of the difference between the estimated treatment effect and the true value. It captures both the bias and the precision of an estimator. It was estimated by taking the square root of the mean of the squared differences between estimated and true treatment effects from the simulated datasets in each scenario.

Relative bias, RMSE, and standard error. In order to compare the performance statistics (bias, RMSE and standard error) between the boosted model and the covariate balancing methods, we used the log-relative values of the statistic defined as $\log(STAT_{CBPS}/STAT_{GBM})$ where for the statistic of interest, $STAT_{CBPS}$ was the estimated value for the covariate balancing method and $STAT_{GBM}$ was the estimated value for the boosted models method. Negative values of the log-relative statistics were indicative of better performance of covariate balancing when compared to boosted models and positive values indicative of dominance by the boosted models.

Measures of model complexity—To assess the impact of the complexity of the data generating structures on the relative performance of the boosted models and the covariate balancing methods, we proposed a one dimensional measure of complexity that summarizes the different simulation parameters and that can be estimated by practitioners on any data. We defined it as the measure of the proportion of variance that could have been explained in a model (R-square) if polynomial models were considered, relative to the linear specification or logistic regression model. For the complexity of the relationship between the outcome and the predictors or in the propensity score model, we used the R-squares from linear models and the pseudo R-squares from logistic regression models^{18;19}, respectively. For a simulated data and a specific strategy, the following steps were used to measure complexity:

1. A model was fit with the strategy predictors as specified in the data generating section linearly entered in the regression and the model R-square (called R_0^2) recorded.
2. Then a second model was fitted with the linear predictors as well as their quadratic and cubic power and all interactions between the continuous predictors included. The model R-square (called R_1^2) was also recorded. This model will potentially capture most non-linearity that may exist in the relationship.

3. The model complexity R_d^2 equals the difference in R-squares

$$R_d^2 = R_1^2 - R_0^2$$

between the two estimates where values close to 0 reveal data models with relatively little complexity that can be well fitted with simple linear combinations and large values reflect very complex models. For the outcome and propensity score models, we denote such complexity indices by R_{dy}^2 and R_{dt}^2 , respectively, and note they will be specific to each dataset generated.

The impact of model complexity on bias, RMSE and standard error were contrasted between the different values of R_{dy}^2 and R_{dt}^2 . (See eAppendix for details.)

Results

Overall performance

We first report in Table 4 the results of simulations with $\tau = 0$ and an additive linear outcome model (outcome 1), a setting without nuisance parameter and similar to simulations reported by^{5,6}. We report the summary of the results for the treatment generation versions A, E and G; the other versions were in similar directions. For covariate balance, for both boosted models and covariate balancing methods across all different treatment generation versions considered from very simple to very complex (in a logit scale), both methods obtain the desired covariate balance with observed standardized mean differences below the commonly used threshold of 0.1 for acceptable balance²⁰. The covariate balancing method tended to produce better balance as measured by the standardized mean difference than the boosted models method. For the boosted models method, the average standardized mean difference ranged from 0.052 to 0.088 while for the covariate balancing method it ranged from 0.005 to 0.032. These results are similar to the ones reported by⁶ and⁵. In contrast, in all the simulations, except for very few cases, the Kolmogorov-Smirnov statistics were slightly smaller for boosted models than covariate balancing, suggesting that the covariate balancing method tended to balance the means of the distributions of the covariates while the boosted models method slightly better balanced the full distributions of the covariates. Also across the board, including instruments or distractors in the propensity score estimation (strategies 2, 5, or 6), resulted in larger Kolmogorov-Smirnov statistics for both estimation methods. The covariate balancing method produced more variable weights in all but one condition (strategy 1 used with treatment generation G) with the biggest differences for models that included instruments and distractors

For estimation of the treatment effect, the covariate balancing method consistently resulted in smaller bias when simple propensity score models were considered but the trend reversed for complex models such as version G. The absolute bias in version A (additive linear propensity score structure) was between 3 times and 5 times higher for the boosted models method than the covariate balancing method, and in version E, it was about between 2 and 3 times higher for the boosted models method. On the other hand, in version G, the bias in the boosted models method was 2.4 to 3.1 times lower with the exception of strategy 4 where it

was only 1.5 times lower. The RMSE performance was similar between both methods. The ratios in RMSE were between 0.9 and 1.1. Even more similar standard errors were observed between the two estimation methods.

The results of the simulation with possible non-linear relationships between outcomes and confounders are reported in Table 5 where we show the summary statistics averaged across all scenarios. Similar to the previous results, the standardized mean differences from the boosted models method were between 3 and 14 times higher than the ones from the covariate balancing method. The Kolmogorov-Smirnov statistic was similar between both estimation methods under the propensity score model version A that has little to no complexity, but in more complex propensity score model setting G, the statistic was between 1.4 and 1.9 times lower in the boosted models method than the covariate balancing method. In Version E, the Kolmogorov-Smirnov statistic was similar in the 3 strategies (2, 5 and 6) where all the covariates used for the generation of the propensity score model were included in the estimation, but for all the other strategies, the statistic was 1.4 to 1.6 times lower with the boosted models method. As in Table 4, the covariate balancing method yields more variable weights, especially in the presence of instruments and distractors. In terms of bias, consistently across all model strategies, the boosted models method produced higher bias for the simple propensity score version A structure (between 3.4 and 4.8 times higher). But for complex models such as version G, the bias in the boosted models method was consistently smaller on the order of 5.6 to 10.4 times. Even in version E, the bias in the boosted models method was lower. For the RMSE, on average, both estimation method produced similar estimates in version A but with more complicated propensity score versions, the RMSE in the boosted models method was consistently lower. In versions E and G, the RMSE in the boosted models method was lower on the order of 1.1 and 1.3 to 1.4 times respectively. Once again very similar standard errors were observed between both estimation methods. A comparison of the linear (Table 4) and the non-linear (Table 5) outcome model setting results also reveals that even if one method performs better in terms of standardized mean difference or Kolmogorov-Smirnov statistic balance, bias and RMSE can still depend the outcome model or the existence of instruments and distractors, and may not always favor the method that yields better balance.

In summary, while the covariate balancing method with only main effects for covariates tended to produce better covariate standardized mean difference balance, the boosted models method produced slightly better Kolmogorov-Smirnov statistic balance especially in complex propensity score settings. For simple outcome and propensity score structures, the covariate balancing method outperformed the boosted models method in bias and both lead to similar RMSE and standard error, but in more complex propensity score settings the boosted models method performed better than the covariate balancing method in both bias and RMSE while the standard error remains similar between the methods. For nearly all conditions, the covariate balancing method yields more variable weights (5% to 49 % greater variability, see Tables 4, 5 and eAppendix) but this does not necessarily translate to larger standard errors for the treatment effect: the standard errors are more similar across both methods than the variance of the weights and more often smaller for the covariate balancing method than the boosted models method. Hence, variance in the weights does not necessarily translate to more variable treatment effect estimates.

Method diagnostic

To provide researchers with ways of assessing circumstances where one method might be more appropriate, we now report the performance statistics by the model complexity indicators defined in section. Results of the estimated bias and RMSE depending on the level of the complexity in the outcome and in the treatment assignment are presented in Figures 1 and 2 where for each outcome complexity R_{dy}^2 and each treatment complexity R_{dt}^2 the average statistic across scenarios was computed and the log-relative statistic ($\log(STAT_{CBPS}/STAT_{GBM})$) presented. The figures for standard errors are reported in the eAppendix. The generalized additive model²¹, a method that combines the advantages of nearest neighbor and kernel methods, was used to create a continuous map surface of the log-relative bias and RMSE by smoothing over the complexities R_{dy}^2 and R_{dt}^2 . Other than smoothing the edges, the smoothed heat plots provided the same information as the original values that had not been smoothed. In the figures, the smoothed log-relative statistics were truncated at most to be between -4 and 4. Values of the log-relative statistic (e.g. for bias) at 1, 2, 3 and 4 suggest that the bias in the covariate balancing method is 2.7, 7.4, 20.1 and 54.6 times larger than the bias in the boosted models method, respectively. The values at -1, -2, -3 and -4 will reflect the opposite with the covariate balancing method producing the smaller bias. For comparisons where the range of difference between the methods is smaller (e.g. RMSE), the range was truncated to values that capture the key differences; for RMSE, we used -2 to 2.

In Figure 1 we present the bias and RMSE plots for strategies 1 and 2 where predictors of both treatment and outcome or instruments were used. For strategy 1 where only predictors of both outcome and treatment were included in the analysis, we observed only a small range in the improvement R_{dt}^2 in the propensity score estimation. In this scenario, both estimation methods performed similarly in RMSE in most cases with the exception of cases with large complexity in the outcome and propensity score model where the boosted models method had lower RMSE. A somewhat different picture was revealed when it comes to bias where, for very low treatment complexity index, the covariate balancing method performed best, but as the propensity score model started getting more complex, the boosted models method outperformed the covariate balancing method. Similar results were observed for modeling strategy 2 (instruments included in the model) where for RMSE, both methods tended to be similar. For bias, the covariate balancing method outperformed the boosted models method whenever there was almost no complexity in the propensity score model and the boosted models method outperformed the covariate balancing method for greater level of complexity. Interestingly, the standard errors were similar between the methods with only differences (on the order of 1.28 times larger or smaller) observed when the complexity of the models is high (see plots in the eAppendix).

In the more common situation of including all available covariates (a model where instruments and distractor variables ended up in the propensity score model), both estimation methods still had comparable RMSE, except when there was relatively large complexity in both outcome and propensity, then the boosted models method performed better (Figure 2a). In terms of bias, the covariate balancing method only outperformed the boosted models method when both outcome and propensity score model complexities were

very low; otherwise, the boosted models method produced treatment effect estimates with lower bias most of the time (Figure 2b).

In general, results were similar under the different propensity score estimation strategies as demonstrated in the plots of the averages across all strategies (Figures 2c and 2d). Taken together these results suggest that when a researcher is confident about the confounders or variables that determine treatment assignment and the complexity in the propensity model is low, the covariate balancing method will likely produce better results as it has RMSEs similar to the ones obtained when using the boosted models method while at the same time enjoying less bias. Conversely, the boosted models method will perform better when complexity in the propensity score model and/or outcome model is high.

Other design-based alternatives to covariate balancing propensity scores

Several other techniques similar to the covariate balancing propensity method that estimate propensity score weights with methods beyond logistic regression have been proposed^{22;23;24}. Many of them design weights that are not equal to the inverse of the propensity score but are chosen explicitly to optimize balance between the covariate distributions in the exposed and control groups. The method of entropy balancing proposed by²², for example, provides such a way of estimating weights for each control observation so that the sample means (or moments) of the selected covariates are identical between the treatment and weighted control groups. We also compared the entropy balancing method to the boosted models and the covariate balancing methods using the same simulation setups in the study design section to assess comparability with other methods (see eAppendix). The performance of entropy balancing mirrors the generalized boosted models method across simulations. When the model complexity in the propensity score is low, entropy balancing and the boosted models method are similar and produce better results than the boosted models method but with high model complexity, the boosted models method outperforms both of them.

Discussion

In a field where the analysis of observational study data is common, propensity score methods have become a useful tool for epidemiologists. In practice, it has been reported that propensity scores that are estimated using logistic regression can have disadvantages²⁵. In particular, such propensity scores often result in lingering imbalances in covariate distributions between the treatment and control samples, leaving open the possibility that claimed associations between the treatment and the outcome can be explained by observed confounders.

Generalized boosted models and covariate balancing propensity scores methods are two advanced propensity score estimation methods that have been found in previous research to circumvent some of the shortcomings of logistic regression. In this paper, we analyze millions of simulated datasets to illuminate for practitioners when they can expect the covariate balancing method to outperform the boosted models method and vice versa. In short, if the parametric assumptions of the logistic regression model are satisfied and the

outcome model is linear in the covariates, the covariate balancing method performs extremely well. On the other hand, when the true propensity score model is substantially nonlinear or contains important interactions in the log-odds scale or the outcome model is not linear in the observed confounders, the flexibility of the boosted models method tends to provide better results. The relative performances of the methods also depend on the kind of covariate balance that matters in the outcome model. As such, researchers might have to think about the outcome model and its complexity when deciding which method to use. In addition to these qualitative guidelines, we provide a data-analytic framework for choosing between the two methods. After selecting potential confounders and avoiding possible distractors and instruments, practitioners can compute our proposed measure of complexity R_d^2 from their data and use results from this study as a guide for the choice between the generalized boosted models and the covariate balancing propensity scores methods. We hope that this guidance will help epidemiologists to choose the best tool for the job when estimating propensity scores. Additionally, we believe our simulation structure will provide a useful framework for future work comparing alternative propensity score methods. Simulations using the method of entropy balancing revealed that it was essentially equivalent in our performance measures to the covariate balancing method. Thus, the comparison shown here between the boosted models and the covariate balancing methods is also representative of a comparison between entropy balancing and the boosted models method (see eAppendix). The field is rapidly developing robust tools for propensity score estimation, each with their own strengths and limitations^{26;27;28;29}. Future work should aim to expand on the assessment of optimal propensity score estimation diagnostics tools, how trimming weights might impact our results in light of the impact of weight trimming reported by³⁰ and understand when these methods might be most useful relative to one another in similar ways to our work here with the generalized boosted models and the covariate balancing propensity scores methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Source: All phases of this study were supported by an NIH grant R01DA034065 from the National Institute on Drug Abuse (NIDA) and a funding from the RAND center for causal inference.

References

1. Field RW, Steck DJ, Smith BJ, Brus CP, Fisher EL, Neuberger JS, Platz CE, Robinson RA, Woolson RF, Lynch CF. Residential radon gas exposure and lung cancer the iowa radon lung cancer study. *American Journal of Epidemiology*. 2000; 151(11):1091–1102. [PubMed: 10873134]
2. Barnoya J, Glantz S. Cardiovascular effects of secondhand smoke: nearly as large as smoking. *Circulation*. 2005; 111:2684–2698. [PubMed: 15911719]
3. Rubin D. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*. 1979; 74:318–324.
4. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:4155.

5. Lee B, Lessler J, Stuart E. Improving propensity score weighting using machine learning. *Statistics in Medicine*. 2009; 29(3):337–346.
6. Wyss R, Ellis R, Brookhart A, Girman C, Jonsson Funk M, LoCasale R, Sturmer T. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American Journal of Epidemiology*. 2014; 180(6):645–655. [PubMed: 25143475]
7. McCaffrey D, Ridgeway G, Morral A. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*. 2004; 9(4):403–425. [PubMed: 15598095]
8. Imai K, Ratkovic M. Covariate balancing propensity score. *Journal of the Royal Statistical Society. Series B (Methodological)*. 2014; 76(1):243–263.
9. Stuart E. Matching methods for causal inference: A review and a look forward. *Statistical Science*. 2010; 25(1):1–21. [PubMed: 20871802]
10. Imbens G. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*. 2004; 86:4–29.
11. Kurth T, Walker A, Glynn R, Chan K, Gaziano J, Berger K, Robins J. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*. 2006; 163:262–270. [PubMed: 16371515]
12. Imai K, King G, Stuart E. Misunderstandings among experimentalists and observationalists in causal inference. *Journal of the Royal Statistical Society. Series A*. 2008; 171:481–502.
13. Hansen L. Large sample properties of generalized method of moments estimators. *Econometrica*. 1982; 50:1029–1054.
14. Owen, A. *Empirical Likelihood*. Boca Raton, FL: Chapman and Hall CRC; 2001.
15. Breiman, L., Friedman, J., Olshen, R., Stone, C. *Classification and Regression Trees*. Boca Raton, FL: Chapman and Hall CRC; 1984.
16. Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., Griffin, B. Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. 2013. [Online]. Available: <https://cran.r-project.org/web/packages/twang/index.html>
17. McCaffrey, D., Burgette, L.F., Griffin, B., Martin, C. Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang sas macros. 2014. [Online]. Available: <http://www.rand.org/pubs/tools/TL136.html>
18. Cox, D., Snell, E. *Analysis of Binary Datas*. Boca Raton, FL: Chapman and Hall CRC; 1989.
19. Maddala, G. *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press; 1983.
20. Austin P. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in medicine*. 2007; 26(16):3078–3094. [PubMed: 17187347]
21. Hastie, T., Tibshirani, R. *Generalized Additive Models*. London: Chapman & Hall; 1990.
22. Hainmueller J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*. 2012; 20(1):25–46.
23. Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*. 2015; 110(511):910–922.
24. Chan KCG, Yam SCP, Zhang Z. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society. Series B (Methodological)*. 2016; 78(3):673–700.
25. Westreich D, Lessler J, Funk M. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*. 2010; 63(8):826–833. [PubMed: 20630332]
26. van der Laan MJ. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*. 2014; 10(1):29–57. [PubMed: 24516006]
27. Hill J. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*. 2010; 20(1):217–240.
28. Liaw A, M W. Classification and regression by random forest. *R News*. 2002; 2(3):18–22.

29. Breiman L. Random forests. *Machine Learning*. 2001; 45:5–32.
30. Lee B, Lessler J, Stuart E. Weight trimming and propensity score weighting. *Plos One*. 2011; 6(3):e18 174.

a. RMSE Strategy 1: Predictors of both Y and T

Heat density scale: -2, 2

b. Bias Strategy 1: Predictors of both Y and T

Heat density scale: -4, 4

c. RMSE Strategy 2: Predictors of T , regardless of Y

Heat density scale: -2, 2

d. Bias Strategy 2: Predictors of T , regardless of Y

Heat density scale: -4, 4

Figure 1.

Simulation results comparing generalized boosted models to covariate balancing propensity scores under different strategies for the selection of the predictors to be included in the analysis. Strategy 1 includes predictors of treatment T and outcome Y while strategy 2 includes predictors of T , regardless of Y . Note: RMSE = Root Mean Square Error.

a. RMSE for all predictors included

Heat density scale: -2, 2

b. Bias for all predictors included

Heat density scale: -4, 4

c. RMSE averaged across all strategies

Heat density scale: -2, 2

d. Bias averaged across all strategies

Heat density scale: -4, 4

Figure 2.

Simulation results comparing generalized boosted models to covariate balancing propensity scores under strategies of including all available covariates and a summary across all strategies. Note: RMSE = Root Mean Square Error.

Table 1

Conjectures on the relative performance of CBPS and GBM across different conditions

Factor	Conditions for which	
	GBM will yield more accurate treatment effects	CBPS will yield more accurate treatment effects
Extraneous variables present	Yes	No
Propensity score complexity	PS model is non-additive or nonlinear	PS model is additive in main effects
Strength of selection	Treatment is strongly related to covariates	Treatment is weakly related to covariates
Conditional mean of the outcomes	Conditional mean is non-additive or nonlinear	Conditional mean is additive in main effects
Residual variance in outcomes	Covariates explain much of the variance in outcomes	Covariates explain little of the variance in outcomes

Note: “accurate” is for both bias and mean squared error.

Table 2

Versions used for generating treatment assignment

A. Additive and linear (main effects terms only):

$$Version_A = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_7 X_7$$

B. Mild non-linearity (1 quadratic term):

$$Version_B = Version_A + \alpha_8 X_2^2$$

C. Moderate non-linearity (3 quadratic terms):

$$Version_C = Version_A + \alpha_9 X_2^2 + \alpha_{10} X_4^2 + \alpha_{11} X_7^2$$

D. Mild non-additivity (4 two-way interaction terms):

$$Version_D = Version_A + \alpha_{12} X_1 X_3 + \alpha_{13} X_2 X_4 + \alpha_{14} X_4 X_5 + \alpha_{15} X_5 X_6$$

E. Mild non-additivity and non-linearity (1 quadratic term and 4 two-way interaction terms):

$$Version_E = Version_D + \alpha_{16} X_2^2$$

F. Moderate non-additivity (9 two-way interaction terms):

$$Version_F = Version_D + \alpha_{17} X_5 X_7 + \alpha_{18} X_1 X_6 + \alpha_{19} X_2 X_3 + \alpha_{20} X_3 X_4 + \alpha_{21} X_4 X_5$$

G. Moderate non-linearity and non-additivity (3 quadratic term and 9 two-way interaction terms):

$$Version_G = Version_F + \alpha_{22} X_2^2 + \alpha_{23} X_4^2 + \alpha_{24} X_7^2$$

Table 3

Conditional mean models for generating outcomes

1. Additive and linear (main effects terms only):
$Y = \delta_0 + \theta T + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \delta_5 X_8 + \delta_6 X_9 + \delta_7 X_{10}$
2. Mild non-linearity (2 individual non-linear variables):
$Y = \delta_0 + \theta T + \delta_1 X_1 + \delta_2 X_2^2 + \delta_3 X_3 + \delta_4 \exp(1.3 X_4) + \delta_5 X_8 + \delta_6 X_9 + \delta_7 X_{10}$
3. Moderate non-linearity (exponential interaction among all confounders):
$Y = \delta_0 + \theta T + \exp(\delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \delta_5 X_8 + \delta_6 X_9 + \delta_7 X_{10})$
4. Severe non-linearity (non-linear, exponential interaction and additive terms):
$Y = \delta_0 + \theta T + \exp(\delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3) + \delta_4 \exp(1.3 X_4) + \delta_5 X_8 + \delta_6 X_9 + \delta_7 X_{10}$
5. Sinusoidal non-linearity (Sine function of confounders):
$Y = \delta_0 + \theta T + 4 \sin(\delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \delta_5 X_8 + \delta_6 X_9 + \delta_7 X_{10})$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Performance of the methods when $\tau = 0$ and linear outcome model considered

Scenarios		GBM						CBPS					
Version	Strategy	100 Bias	100 SD	100 RMSE	IPW SD	ASAMD	KS	100 Bias	100 SD	100 RMSE	IPW SD	ASAMD	KS
A: additive linear	1	1.26	6.77	6.95	0.43	0.055	0.051	0.26	6.69	6.70	0.51	0.005	0.056
	2	1.59	7.48	7.73	0.55	0.087	0.081	0.43	8.00	8.02	0.88	0.018	0.072
	3	1.35	6.49	6.72	0.41	0.052	0.056	0.32	6.33	6.35	0.51	0.006	0.059
	4	1.21	6.81	6.99	0.47	0.058	0.057	0.38	6.79	6.80	0.61	0.008	0.056
	5	1.69	7.23	7.53	0.52	0.080	0.084	0.48	7.58	7.61	0.86	0.019	0.074
	6	2.02	7.11	7.52	0.48	0.074	0.090	0.66	7.51	7.56	0.84	0.021	0.079
E: mild non-additive non-linear	1	0.88	6.91	7.00	0.47	0.059	0.052	0.34	6.75	6.77	0.51	0.006	0.086
	2	0.79	7.64	7.71	0.57	0.088	0.085	0.24	8.14	8.15	0.86	0.017	0.091
	3	0.77	6.62	6.71	0.45	0.055	0.057	0.22	6.37	6.38	0.51	0.007	0.086
	4	0.58	7.04	7.08	0.52	0.058	0.058	0.27	6.97	6.98	0.64	0.009	0.079
	5	0.75	7.37	7.43	0.54	0.082	0.089	0.21	7.65	7.66	0.84	0.018	0.091
	6	0.96	7.27	7.37	0.50	0.076	0.094	0.35	7.59	7.60	0.82	0.020	0.092
G: moderate non-additive non-linear	1	0.25	7.32	7.33	0.62	0.060	0.054	0.73	6.84	6.90	0.62	0.016	0.105
	2	0.23	7.96	7.97	0.63	0.080	0.117	0.72	7.20	7.25	0.71	0.032	0.172
	3	0.20	7.02	7.03	0.58	0.056	0.061	0.55	6.49	6.53	0.62	0.014	0.103
	4	0.30	7.41	7.41	0.64	0.057	0.059	0.45	6.91	6.93	0.70	0.015	0.105
	5	0.24	7.67	7.68	0.59	0.073	0.120	0.57	6.87	6.90	0.71	0.028	0.173
	6	0.23	7.49	7.50	0.54	0.068	0.123	0.56	6.83	6.88	0.71	0.024	0.173

Note: 100Bias = 100× absolute value of Bias, 100RMSE = $100 \times \text{Root Mean Square Error (RMSE)}$, IPW SD = Inverse-Probability Weights Standard Deviation, ASAMD = Average Standardized Absolute Mean Difference, KS = Kolmogorov-Smirnov Statistic.

Table 5

Average Performance of GBM and CBPS propensity score methods across all scenarios

Scenarios		GBM						CBPS					
Version	Strategy	100 Bias	100 SD	100 RMSE	IPW SD	ASAMD	KS	100 Bias	100 SD	100 RMSE	IPW SD	ASAMD	KS
A: additive linear	1	1.09	6.35	6.51	0.398	0.050	0.048	0.23	6.41	6.41	0.442	0.004	0.054
	2	1.35	6.91	7.12	0.497	0.077	0.072	0.34	7.33	7.34	0.741	0.013	0.066
	3	1.24	6.10	6.32	0.378	0.048	0.053	0.26	6.12	6.13	0.448	0.004	0.057
	4	1.16	6.33	6.51	0.425	0.053	0.052	0.26	6.43	6.44	0.531	0.006	0.054
	5	1.47	6.67	6.94	0.473	0.072	0.076	0.40	6.98	7.00	0.733	0.015	0.069
	6	1.67	6.63	6.97	0.440	0.067	0.081	0.50	6.96	6.99	0.721	0.016	0.073
E: mild non-additive non-linear	1	0.60	6.49	6.54	0.436	0.054	0.048	2.15	6.52	7.12	0.452	0.005	0.079
	2	0.66	7.11	7.17	0.520	0.079	0.076	2.27	7.57	8.16	0.739	0.013	0.083
	3	0.67	6.21	6.28	0.415	0.051	0.054	2.16	6.21	6.86	0.457	0.005	0.079
	4	0.53	6.52	6.57	0.471	0.054	0.053	2.08	6.62	7.19	0.562	0.007	0.073
	5	0.77	6.83	6.91	0.495	0.074	0.079	2.28	7.17	7.80	0.730	0.014	0.083
	6	0.99	6.77	6.91	0.461	0.069	0.085	2.29	7.14	7.77	0.716	0.015	0.084
G: moderate non-additive non-linear	1	0.47	6.75	6.79	0.563	0.054	0.050	4.95	6.62	9.04	0.548	0.012	0.096
	2	0.64	7.32	7.37	0.577	0.072	0.104	5.12	6.95	9.42	0.630	0.027	0.156
	3	0.55	6.47	6.52	0.526	0.051	0.056	4.96	6.34	8.86	0.554	0.011	0.094
	4	0.52	6.73	6.77	0.578	0.051	0.053	5.04	6.63	9.14	0.622	0.012	0.096
	5	0.75	7.02	7.10	0.543	0.067	0.107	5.14	6.66	9.23	0.633	0.023	0.156
	6	0.91	6.93	7.04	0.499	0.063	0.110	5.09	6.68	9.21	0.636	0.021	0.157

Note: 100Bias = 100× absolute value of Bias, 100RMSE = $100 \times \text{Root Mean Square Error (RMSE)}$, IPW SD = Inverse-Probability Weights Standard Deviation, ASAMD = Average Standardized Absolute Mean Difference, KS = Kolmogorov-Smirnov Statistic.