

RESEARCH ARTICLE

Should a propensity score model be super? The utility of ensemble procedures for causal adjustment

Shomoita Alam¹ | Erica E. M. Moodie¹  | David A. Stephens²

¹Department of Epidemiology,
Biostatistics and Occupational Health,
McGill University, Montreal, Canada

²Department of Mathematics and
Statistics, McGill University, Montreal,
Canada

Correspondence

Erica E. M. Moodie, Department of
Epidemiology, Biostatistics and
Occupational Health, McGill University,
Montreal, Canada.

Email: erica.moodie@mcgill.ca

Funding information

Natural Sciences and Engineering
Research Council of Canada; Fonds de
Recherche du Québec - Santé,
Grant/Award Number:
Chercheur-Boursier Career award

In investigations of the effect of treatment on outcome, the propensity score is a tool to eliminate imbalance in the distribution of confounding variables between treatment groups. Recent work has suggested that Super Learner, an ensemble method, outperforms logistic regression in nonlinear settings; however, experience with real-data analyses tends to show overfitting of the propensity score model using this approach. We investigated a wide range of simulated settings of varying complexities including simulations based on real data to compare the performances of logistic regression, generalized boosted models, and Super Learner in providing balance and for estimating the average treatment effect via propensity score regression, propensity score matching, and inverse probability of treatment weighting. We found that Super Learner and logistic regression are comparable in terms of covariate balance, bias, and mean squared error (MSE); however, Super Learner is computationally very expensive thus leaving no clear advantage to the more complex approach. Propensity scores estimated by generalized boosted models were inferior to the other two estimation approaches. We also found that propensity score regression adjustment was superior to either matching or inverse weighting when the form of the dependence on the treatment on the outcome is correctly specified.

KEYWORDS

average treatment effect, confounding, covariate balance, inverse probability weighting, matching, propensity score, Super Learner

1 | INTRODUCTION

Randomized control trials are considered to be the gold standard to estimate the causal effects of a treatment (or more generally, an exposure) on an outcome, as their design automatically controls for confounding of the relationship between treatment and outcome by other covariates, observed or otherwise. In the standard causal language, a confounder is a variable that is a cause of both treatment and outcome, such that there is an association or dependence between the confounding variable and the treatment and outcome. Randomization breaks this dependence and leads to *balance*; essentially, this corresponds to independence of the possible confounding predictors and the assigned treatment. Balance on confounding variables is a sufficient condition to allow consistent estimation of causal effects using simple two-group comparisons. Observational (nonrandomized) studies can also be used to make causal inferences; however, such studies are typically not balanced, in that the baseline characteristics may be systematically different in the treatment and the control groups, leading to confounding of the effect of treatment on the outcome. Therefore, adjustment techniques aim to induce covariate balance in the treatment and the control groups in order to obtain valid estimates of causal treatment effects.

The *propensity score*¹ can be used to eliminate imbalance in the distribution of the covariates, as within strata of the observed propensity scores, the treated and untreated groups are indistinguishable in terms of their covariate profile. The propensity score may be used as a covariate in a regression model of the outcome and treatment, for matching of the treated and untreated subjects,² or for reweighting samples according to the inverse of their probability of being treated to estimate an average treatment effect (ATE). Simple parametric modeling approaches such as logistic regression are widely used to estimate the propensity score,³ and Rubin⁴ suggests that complex models may include interaction and/or quadratic terms. The functional relationship between the treatment and other covariates involved in estimating propensity score using a parametric model must be correctly specified; if this is violated, it may result in covariate imbalance as well as inefficient estimation of the treatment effect.⁵

As an alternative to simple parametric approaches, machine learning and data-adaptive methods such as classification trees, neural networks, and boosting methods have been suggested; see, for example other works.^{6–9} Combinations of flexible machine learning methods, termed *ensemble* methods, have also been proposed. Perhaps, most prominently, van der Laan et al¹⁰ proposed an ensemble method called Super Learner that can use both parametric and non-parametric techniques in a data-adaptive fashion. The Super Learner model averaging strategy involves computing cross-validated estimates of the empirical predictive error of a number of candidate models based on a selected loss function; a Super Learner estimator is then formed using a weighted linear combination of the candidate predictive models. Pirracchio et al¹¹ considered Super Learner as a propensity score fitting approach and concluded that it improves covariate balance and reduces bias even in case of serious model misspecification. However, overall, the improvements in estimation were modest. Diaz and Kelly¹² focused on inverse weighted estimators and suggested that, in the absence of subject matter knowledge regarding parametric functional forms of the propensity score, predictive accuracy should be used to select an estimator among a collection of candidates. These authors also advocated the use of Super Learner, as have others.¹³ Super Learner has also been considered in the context of longitudinal data, where it has been found useful in the presence of model misspecification.¹⁴ Overall, Super Learner is now being widely adopted in the causal inference literature and in applications.^{15–18}

In much of this literature, and certainly in most applications, adjustment via the propensity score is achieved in a *singly robust* fashion, that is, designed to give consistent inference on the causal estimand under an assumption of correct specification (at least within a class of parametric, flexible, or ensemble procedures). *Doubly robust* procedures that provide consistent estimation if either the propensity score or a proposed conditional outcome mean model—as would be utilized in standard regression—are correctly specified are also well established in the statistical literature, but such procedures are not as widely adopted in practice. In any case, the representation of the propensity score is an important component of all of these adjustment procedures, and in this paper, we focus on this component.

The propensity score literature incorporating Super Learner speaks compellingly to its use, but simulation studies presented as supporting evidence are often based on simple settings that—even when considering higher dimensional covariates—may fail to capture the complexity of real-world analyses. In practice, Super Learner has been found to induce positivity violations and reduced covariate balance.¹⁹ Further, the benefits found are often modest, particularly relative to the computational burden of ensemble approaches that often render in-depth simulation studies prohibitive, as variance assessments via bootstrapping are infeasible for large-scale investigations. Motivated by the inconsistency in the simulation and real-data findings, we undertook a comprehensive assessment of balance, bias, and efficiency for a range of propensity score modeling approaches focusing predominantly on simulations drawing from real data but also considering 21 scenarios, covering most of the cases explored in previous papers, as well as several new scenarios. We investigate the performance of logistic regression using main effects and interaction terms, Super Learner and generalized boosted methods for propensity score modeling in estimating the ATE using propensity score regression, propensity score matching and inverse probability of treatment weighting (IPTW).

In the following section, we summarize desirable properties of a propensity score and the assumptions needed to estimate a causal quantity effectively. In Section 3, we describe our simulation approach, focusing primarily on the real data used for our simulations. Section 4 examines the results of the real-data simulations in detail and summarizes also the simulations based on entirely synthetic data. Section 5 concludes with a discussion of the findings and their implications for real-data analyses.

2 | REQUIRED PROPERTIES IN PROPENSITY SCORE CONSTRUCTION

We now recap the required or desirable properties of a method proposed for constructing a propensity score. We denote the response or outcome by Y , the exposure or treatment by A , measured confounding variables by W , unmeasured confounding variables by U . Other covariates may also be recorded: we denote by X covariates that predict (or are correlated

with) outcome only, by Z covariates that predict (or are correlated with) treatment only, and by V covariates that are unrelated to both treatment and outcome; X are termed pure predictors of outcome, Z are termed instruments or instrumental variables, and V are called spurious variables. Note that unmeasured variables would be classified as X , Z , or V if measured can be ignored in the current context.

In the binary exposure context, a propensity score model is a model for the probability of exposure (say, $A = 1$) given other measured quantities (W, Z, X, V). The following results hold.

1. **Model construction:** The propensity score model must induce balance on confounding variables but does not need to induce balance on any other variables. That is, we should construct the model $\Pr[A = 1|W]$ and define the propensity score as

$$g(W) = \Pr[A = 1|W]. \quad (1)$$

This may be achieved by regressing A on W in the observed data, and omitting instruments Z .^{3,20,21}

2. **Balance:** A propensity score correctly specified according to (1) induces a balancing mechanism in that data with identical observed values of $g(W)$ can be compared directly in terms of their outcome. In order to assess the value of a proposed propensity score model for adjustment, it is imperative to check for balance in the observed data within strata of the propensity score.
3. **Efficiency:** When considering the causal estimator constructed using any of the standard methods (regression, weighting or matching), potential competing models based on (W, Z) or (W, V) , that is,

$$\Pr[A = 1|W, Z] \quad \Pr[A = 1|W, V]$$

will be less efficient for estimating the causal estimand. The critical relevant issue is that construction of a propensity score model is **not** an exercise in best prediction of A given the available information but, instead, is an exercise in inducing balance.

4. **Positivity:** A propensity score model that predicts treatment **precisely** leads to a violation of *positivity* (or the experimental treatment assignment assumption) that states that causal comparisons can only be carried out between truly comparable groups in which there is a positive probability of both receiving and not receiving treatment.
5. **Pure predictors of outcome:** Variables X do not need to be included in the propensity score as they are unrelated to A . However, the propensity score model based on $\Pr[A = 1|W, X]$ still induces balance on the confounders W and does not improve predictive capability in terms of predicting exposure. However, including X in the propensity score model may—for example, in the case of propensity score regression—reduce the variance of the causal estimator by dint of its association with the outcome Y .
6. **Unmeasured confounding:** If there exist unmeasured confounders U , then only the “oracle” model $\Pr[A = 1|W, U]$ would yield balance in the required fashion to produce consistent estimation. Adjustments based on $\Pr[A = 1|W]$ will not in general yield consistent estimators.

These considerations are only immediately relevant in the hypothetical setting where the researcher can identify each of the components (W, Z, X, V) in their data set; in practice, the variables do not come conveniently labeled in this way. Nevertheless, the considerations do lay out the principles for optimal construction of the propensity score model. Furthermore, it might also be that, in the presence of unmeasured confounding, the classification of a variable would change, or that using an apparently spurious predictor would introduce confounding through a backdoor path.²² In this paper, we proceed as if all confounders are measured, and we focus on procedures for incorporating them optimally.

As mentioned in point 2. above, a key issue in the use of the propensity score adjustment is correct specification; we require that the model in (1) is correctly specified in order for adjustments based on the propensity score to be effective. The correct specification involves both the W variables included and the functional form of the proposed model for $\Pr[A = 1|W]$. In simpler approaches, parametric models $\Pr[A = 1|W] \equiv \Pr[A = 1|W; \beta]$ are proposed; in methods based on ensemble approaches, the specification is “algorithmic” in nature. The main theoretical advantage of ensemble methods is that they are more flexible than standard parametric approaches and, therefore, can consistently estimate $\Pr[A = 1|W]$ for a broader class of data generating mechanisms. However, as we attempt to reiterate throughout this paper, consistent estimation of $\Pr[A = 1|W]$ is not the primary goal of any propensity score-based adjustment, and it may be that much simpler models can achieve adequate adjustment.

3 | SIMULATION APPROACH

The complexities of real-data structures can provide different insights relative to completely artificial data and may be more reflective of the “real-world” performance of a method. We adopt this approach, sometimes known as *plasmode* simulation²³ in our study, as it represents a complement to the highly controlled and simplified settings most commonly adopted. Specifically, motivated by discrepancies between what we have observed in the literature and our experience in application, our primary simulations were based on real data, with all the inherent complexities, correlations, and nonlinearities that may naturally exist remaining unknown to the analyst. We also performed an extensive series of simulations using entirely synthetic data, based on models previously considered in the literature but spread across multiple papers. Our simulations compare the performance of several forms of propensity score estimation (logistic regression, generalized boosted models, and Super Learner) and several utilizations of the propensity score (regression, matching, weighting).

3.1 | Simulation study derived from real-data structure

The basis of our simulation studies is a real data set in which the relationship between the variables (outcome, treatment, and other covariates) has a considerable complexity. The First Steps program, which was designed to evaluate and monitor programs and services for low income and other high risk women and children in Washington State, United States, began in August 1989. For the purposes of our analyses, the units of observation are mother-infant pairs, and the binary treatment variable is an indicator of whether the mother participated in the First Steps program. Covariates retained for use in our simulation were mother's age, race (white, asian, or other), parity (number of previous pregnancies, coded 0, 1, or >1), marital status, smoking status, weight prior to pregnancy, and education level, as well as the child's sex.

In the real data setting, the true impact of the treatment on outcome, and the potentially confounding relationships between variables, are unknown. However, for a comparative study of the effectiveness of different adjustment approaches, we may retain the observed treatment and covariate data and define a (structural) outcome model that defines the magnitude of the effect of these variables on outcome. This approach does not define the confounding structure (that is, the model $\Pr[A = 1|W]$ from (1) is not specified in the simulation), but it does define the causal parameter of interest in the outcome model, which may then be treated as the true value of this quantity. Four data generating scenarios were considered. The exposure variable (A , with $A = 1$ corresponding to participation in the First Steps program) was used as naturally given in the data. In two scenarios, the variables sex, race, parity, and smoking status were set to be confounders (W), in that they appeared to be associated with the exposure, and were used to predict the synthetic outcome variable. The remaining variables were unrelated to the outcome and thus, these were either instruments or spurious variables. Two further scenarios were considered, in which all covariates were used to generate the outcome variable, and hence, all were potential confounders.

The general form of the outcome model used in our simulations, from which the artificial outcome data were generated, supposed an additive treatment effect, that is, with linear predictor

$$\theta\alpha + \mathbf{w}\alpha, \quad (2)$$

where $\mathbf{w}$ represents the row vector of confounders; specifically, there was no interaction between confounders and treatment in the simulation. This structure was chosen for simplicity but could easily be extended to allow for interactions to be present. The continuous outcome of interest was birth weight (Y) of the infant. The data generating parameters were guided by associations observed in the First Steps data between the outcome and the chosen confounders (sex, race, parity, and smoking status); all other variables were set to have no effect on the outcome. The ATE was set to be either null (0 g) or 150 g. Variability was added to the outcome in the form of a random mean-zero Normal variable with variance that was the residual variance found in an analysis of the First Steps data (see Online Supplementary Material). Only the “true” confounders (sex, race, parity, and smoking status) were included. Thus, the four scenarios considered were as follows: (i) four confounders, null ATE; (ii) four confounders, ATE of 150 g; (iii) eight confounders, null ATE; or (iv) eight confounders, ATE of 150 g.

The First Steps database comprises 2500 observations. In our simulations, we considered 500 replicates of simulated samples of size 100, 300, and 500 to assess the performance of the estimators of interest. These simulations drew samples of covariate and exposure variables from the full database with replacement to achieve the desired sample size.

3.2 | Synthetic data simulation

We additionally considered 21 simulation scenarios.

- Four simulation scenarios were reproduced from the work of Pirracchio et al,¹¹ and an additional four were similar to those but considered alternative parameter settings in the propensity score model.
- A further eight simulations were considered using the same data generating scenarios but in which the analyses included instruments in the propensity score models.
- Another simulation scenario was implemented similar to the setting proposed by Kang and Schafer,⁵ which was also discussed in the work of Diaz and Kelly.¹²
- Finally, we considered four new data generating scenarios. These simulations were designed to consider a range of settings in which covariates enter the propensity score model linearly or nonlinearly, so that model misspecification of varying degrees occurs in most of the scenarios.

Full details of these 21 data generating and model fitting settings are provided in the Online Supplementary Materials.

3.3 | Propensity score estimation

In all simulations, whether the data were generated using real covariates from First Steps data set or entirely synthetic data, we fit the propensity score using four approaches:

- logistic regression, with covariates entered only as main effects (PS-LR in Figures and Tables);
- logistic regression, with covariates entered as main effects and all two-way interactions included (PS-LR2);
- generalized boosted models (PS-GBM) as suggested by Kang and Schafers⁵;
- the ensemble approach Super Learner (PS-SL).

The generalized boosted models were implemented using the `twang` package in R.²⁴ The Super Learner was implemented using the same libraries employed by Pirracchio et al,¹¹ with default settings for all methods and the cross-validated L2 squared error as the loss function. Support vector machines were excluded from the Super Learner ensemble for the real (First Steps) data simulations analysis due to lack of convergence but were used in the simulations based on entirely synthetic data. Standard errors and coverage of 95% Wald-type confidence intervals were computed using a nonparametric bootstrap with 100 resamples to estimate the standard error. Due to the computational burden of computing bootstrapped standard errors for the ensemble approach, cross-validation was limited to five-fold. While the bootstrap is known not to be valid for matching procedures,²⁵ the `Matching` package in R does not currently provide standard errors that account both for the matching and the estimation of the propensity score.

3.4 | Estimation of the ATE

The population ATE is defined as the contrast between the expected outcome had the entire population been treated, relative to the expected outcome had treatment been withheld from the entire population. Several forms of estimation of the ATE were considered, as follows:

- a *naive model*: regressing the outcome on the exposure only;
- a *covariate-adjusted model*: regressing the outcome on all covariates;
- a *propensity score-adjusted model*: regressing the outcome on the exposure and the propensity score (as a linear, main effect term);
- a *matched analysis*: regressing the outcome on the exposure only, in a sample formed by pair-matching based on the propensity score;
- a *weighted analysis*: regressing the outcome on the exposure only, in a sample weighted by the inverse probability of exposure.²⁶

The matched analysis was carried out using the `Matching` package,²⁷ with each subject in the sample is matched once to the nearest subject in the alternative exposure group based on calipers of width of 0.2 of the standard deviation of the logit of the propensity score. Matching was performed with replacement. The latter three of the approaches were employed with each of the four propensity score estimation methods described in the previous section.

In all cases, the covariate-adjusted outcome model was fit as a linear model. Whenever logistic regression was employed to estimate the propensity score, covariates were entered also as linear terms (in some cases with interactions between covariates). In several settings, the true covariate-outcome relationship was in fact linear, and so, we would expect superior performance from covariate adjustment; however, in many settings, only transformed versions of the covariates were available to the analyst so that none of the models could be specified precisely.

The naive model is expected to provide a biased estimator in the presence of confounding, while the covariate adjusted model will provide an unbiased estimator of the ATE when the outcome model is correctly specified. The latter three approaches will provide consistent estimators under correct specification of the propensity score model.

3.5 | Performance metrics

The performance of four approaches to propensity score estimation for each analysis were compared using the following metrics:

- the bias, empirical standard error, and root mean squared error (rMSE) of the estimated ATE;
- the balance between the covariates between exposed and unexposed subjects,²⁸ as assessed by the average standardized absolute mean difference (ASAM), averaged over all covariates and expressed as a percentage;
- the number of subjects discarded by the matching procedure for the matching estimator.

Typically, a standardized mean difference of 10% or greater is considered to be indicative of poor covariate balance. The computational time required to calculate bootstrap standard error from a *single* simulated sample was 20 to 30 minutes, depending on the computational capacities of the nodes used in the Compute Canada clusters used for the simulation study. Thus, due to the extensive number of simulations and the significant computational burden of the generalized boosted models and the ensemble approach, bootstrapping was limited to the null ATE in the real (First Steps) data simulations.

We also report the distribution of the inverse probability weights (Online Supplementary Materials) and the predictive performance (Appendix) of each propensity score estimation method, as measured by the area under the receiver operating curve (AUC)²⁹ and the predictive accuracy, defined as

$$\frac{1}{n} \{ \mathbb{1}(\text{propensity score} \geq 0.5 \ \& \ A = 1) + \mathbb{1}(\text{propensity score} < 0.5 \ \& \ A = 0) \}.$$

All analyses were performed using R statistical software 3.3.1 using high performance computing resources of Compute Canada.³⁰ R code for the First Steps-based simulation is provided in the Online Supplementary Materials.

4 | RESULTS

4.1 | Results: real-data simulations

Results of the real-data simulations with a null ATE are presented in Tables 1 and 2; the nonnull ATE results are similar and are provided in the Online Supplementary Materials.

As expected, bias, and MSE were lowest in the (correctly specified) covariate adjusted model. Across analytic approaches, for a given propensity score estimation approach, propensity score regression outperformed both matching and inverse weighting, supporting findings from other studies that have shown direct adjustment to be superior to either IPTW or matching.³¹ Neither matching nor inverse weighting consistently dominated the other. Examining the results for each given analytic approach, the main effects only logistic regression propensity score is often superior to or competitive with the more complex models, both in terms of bias and variability. While including interactions or other more complex forms sometimes improved balance, the inclusion of all two-way interactions in the propensity score for small samples led to disastrous results, with very few matches and extreme weights (results not shown). For most sample sizes, covariate balance was not achieved for any analytic approach; inverse weighting tended to confer better balance than matching, while there are no explicit measures of balance for outcome regression modeling approaches.

An important issue that arose in our analyses was one of models failing to converge in smaller sample sizes, occasionally due to insufficient numbers (too few individuals in the resampled data with a given characteristic in a regression analysis), but more commonly due to an inability to find sufficient matches for the matched analyses or due to extreme weights in inverse weighting. This issue was most apparent in the bootstrap standard errors at the smallest sample sizes. The

TABLE 1 Simulation results for the simulation using first steps data with a null treatment effect, with only confounders included in the estimated propensity score and outcome models. The ASAM is computed over the confounding variables only

	Absolute Bias	Empirical SE	Bootstrap SE	BS Coverage	rMSE	ASAM(%) ^a	Discarded
Sample Size = 100							
Naive	28.5	170.6	154.1	91.2	173.0	41.5	
Adjusted by Ws	2.3	160.4	160.9	93.6	160.4	41.5	
Adjusted by PS-LR	3.2	162.3	164.6	94.4	162.4	41.5	
Adjusted by PS-LR2	8.1	179.9	197.7	95.4	180.1	41.5	
Adjusted by PS-SL	20.7	163.8	162.1	93.8	165.1	41.5	
Adjusted by PS-GBM	11.4	169.8	179.5	95.8	170.2	41.5	
Logit PS-LR Matching	10.2	220.2	212.8	96.0	220.4	29.4	10.3
Logit PS-LR2 Matching	10.1	212.5	205.6	93.6	212.7	28.3	31.8
SL Matching	23.1	194.5	203.6	97.0	195.9	52.2	9.1
GBM Matching	13.3	223.3	206.0	95.2	223.7	31.9	15.9
Logit PS-LR IPTW	8.6	196.6	196.3	94.0	196.8	18.0	
Logit PS-LR2 IPTW	16.2	201.6	169.9	90.6	202.3	23.6	
SL IPTW	10.0	302.7	324.9	100.0	302.8	101.0	
GBM IPTW	13.8	187.3	162.6	92.2	187.8	20.3	
Sample Size = 300							
Naive	37.9	94.2	89.7	91.6	101.6	32.4	
Adjusted by Ws	4.8	91.7	91.0	94.4	91.8	32.4	
Adjusted by PS-LR	5.3	91.6	91.1	94.2	91.8	32.4	
Adjusted by PS-LR2	3.8	95.2	96.5	94.2	95.3	32.4	
Adjusted by PS-SL	22.7	91.3	89.9	93.8	94.1	32.4	
Adjusted by PS-GBM	1.2	94.7	94.9	93.8	94.7	32.4	
Logit PS-LR Matching	4.5	116.0	113.1	94.8	116.1	18.2	4.1
Logit PS-LR2 Matching	2.6	121.0	126.1	95.2	121.0	17.4	15.3
SL Matching	13.6	120.1	120.8	97.2	120.9	27.1	6.7
GBM Matching	3.3	119.5	121.6	96.0	119.6	18.4	7.8
Logit PS-LR IPTW	2.1	104.0	104.0	93.8	104.0	6.7	
Logit PS-LR2 IPTW	0.1	114.1	119.8	95.4	114.1	8.7	
SL IPTW	16.5	105.8	139.9	98.2	107.1	15.4	
GBM IPTW	2.0	105.4	99.4	92.2	105.4	5.0	
Sample Size = 500							
Naive	35.4	70.3	68.6	92.0	78.7	30.5	
Adjusted by Ws	1.7	69.1	69.7	94.6	69.2	30.5	
Adjusted by PS-LR	1.9	68.8	69.8	94.2	68.8	30.5	
Adjusted by PS-LR2	2.0	70.6	72.2	94.6	70.7	30.5	
Adjusted by PS-SL	14.6	68.3	69.1	95.0	69.9	30.5	
Adjusted by PS-GBM	0.4	70.8	72.1	93.8	70.8	30.5	
Logit PS-LR Matching	1.3	82.3	84.5	95.4	82.3	14.0	3.6
Logit PS-LR2 Matching	2.6	85.9	89.4	96.0	85.9	14.2	11.5
SL Matching	8.0	86.3	92.0	97.6	86.6	21.3	5.2
GBM Matching	0.3	85.6	89.1	95.4	85.6	14.6	6.9
Logit PS-LR IPTW	0.7	74.5	77.1	94.8	74.5	4.7	
Logit PS-LR2 IPTW	2.0	82.3	88.6	96.0	82.4	4.8	
SL IPTW	10.1	76.4	89.5	96.0	77.1	10.8	
GBM IPTW	1.9	77.3	77.5	95.0	77.3	3.1	

^aBalance for covariates child's sex, mother's race, parity, and smoking status.

Abbreviations: ASAM, average standardized absolute mean difference; GBM, generalized boosted model; IPTW, inverse probability of treatment weighting; rMSE, root mean squared error; SL, Super Learner.

most flexible or highly parametric approaches—logistic regression with all pairwise interactions, Super Learner, and the generalized boosted models—were most prone to this problem, which persisted even to the largest sample size of 500 for Super Learner. Full details are provided in the Online Supplementary Materials. Further, while bootstrapping is known to be invalid for matching, performance was good, with the estimated standard errors close to the empirical standard errors and coverage near the nominal level even in small sample sizes.

TABLE 2 Simulation results for the simulation using first steps data with a null treatment effect, with confounders and outcome predictors included in the propensity score and outcome model. The ASAM is computed over the confounding variables only

	Absolute Bias	Empirical SE	Bootstrap SE	BS Coverage	rMSE	ASAM(%) ^a	Discarded
Sample Size = 100							
Naive	46.4	186.9	158.9	89.0	192.6	56.3	
Adjusted by Ws	0.7	178.7	175.5	94.2	178.7	56.3	
Adjusted by PS-LR	2.9	181.8	197.4	96.6	181.8	56.3	
Adjusted by PS-SL	28.8	177.4	197.1	95.4	179.7	56.3	
Adjusted by PS-GBM	38.9	255.5	398.8	99.2	258.5	56.3	
Logit PS-LR Matching	4.8	293.6	318.2	99.4	293.6	35.6	21.1
SL Matching	22.5	238.7	299.6	99.8	239.7	54.3	11.3
GBM Matching	29.6	339.8	414.9	100.0	341.1	53.4	64.8
Logit PS-LR IPTW	4.1	270.0	254.2	95.6	270.0	32.1	
SL IPTW	27.7	339.6	354.6	100.0	340.7	71.5	
GBM IPTW	22.0	190.7	150.9	88.0	191.9	37.4	
Sample Size = 300							
Naive	61.2	99.3	91.1	88.4	116.6	49.6	
Adjusted by Ws	3.2	97.2	94.9	93.4	97.2	49.6	
Adjusted by PS-LR	4.6	98.5	95.8	93.0	98.6	49.6	
Adjusted by PS-LR2	0.4	125.9	166.8	99.4	125.9	49.6	
Adjusted by PS-SL	21.2	97.4	105.3	94.2	99.7	49.6	
Adjusted by PS-GBM	28.7	126.4	131.4	93.8	129.6	49.6	
Logit PS-LR Matching	13.1	176.2	165.8	97.8	176.7	18.9	5.2
Logit PS-LR2 Matching	4.1	271.4	276.8	100.0	271.5	36.5	19.6
SL Matching	7.5	157.6	223.5	99.6	157.8	26.1	4.8
GBM Matching	18.7	252.3	245.6	98.2	253.0	31.9	33.6
Logit PS-LR IPTW	5.0	142.6	135.0	92.8	142.7	13.8	
Logit PS-LR2 IPTW	2.9	254.6	312.7	100.0	254.6	33.2	
SL IPTW	14.3	188.2	304.2	100.0	188.8	23.8	
GBM IPTW	18.8	124.9	99.3	89.0	126.3	20.6	
Sample Size = 500							
Naive	57.7	74.5	69.4	83.4	94.2	48.3	
Adjusted by Ws	3.1	73.9	72.6	93.4	74.0	48.3	
Adjusted by PS-LR	5.8	73.4	73.0	94.2	73.6	48.3	
Adjusted by PS-LR2	3.1	82.4	95.0	95.6	82.5	48.3	
Adjusted by PS-SL	11.2	76.9	83.2	94.4	77.7	48.3	
Adjusted by PS-GBM	24.7	93.0	93.3	92.6	96.2	48.3	
Logit PS-LR Matching	19.8	128.3	126.4	98.0	129.8	14.9	2.6
Logit PS-LR2 Matching	11.2	179.8	223.7	99.6	180.1	21.4	9.3
SL Matching	9.9	130.2	187.4	99.6	130.6	19.2	2.6
GBM Matching	11.8	197.2	196.7	98.2	197.6	26.4	18.4
Logit PS-LR IPTW	12.2	103.3	101.0	94.4	104.1	9.9	
Logit PS-LR2 IPTW	1.0	211.0	257.0	99.8	211.0	27.1	
SL IPTW	13.5	186.3	286.2	100.0	186.8	22.3	
GBM IPTW	21.8	98.8	83.3	92.0	101.1	15.9	

^aBalance for covariates child's sex, mother's age, race, parity, marital status, smoking status, weight prior to pregnancy and education level.

Abbreviations: ASAM, average standardized absolute mean difference; GBM, generalized boosted model; IPTW, inverse probability of treatment weighting; rMSE, root mean squared error; SL, Super Learner.

In summary, for the real-data simulation results in Tables 1 and 2, there is little evidence that the ensemble or other machine learning methods for constructing the propensity score improve estimation of the ATE; indeed, in most cases, the estimator bias and the MSE are typically larger for those methods than for more elementary approaches, albeit often within the Monte Carlo uncertainty tolerance. There may be a small advantage in terms of coverage probabilities, which are sometimes closer to the nominal level; however, for the ensemble methods, the variance and coverage calculations can only be facilitated using the bootstrap, which involves a huge computational overhead. In Table 2, we note that some of the bootstrap-computed sampling variances are extremely large for the smaller sample sizes, illustrating a potential lack of stability of the bootstrap computation in this setting. Note that measures of model fit such as area under the curve (AUC) and predictive accuracy are somewhat similar and generally lower for both logistic regression and Super Learner, despite their rather different performance in terms of MSE for the parameter of interest (see Figures A1- A2).

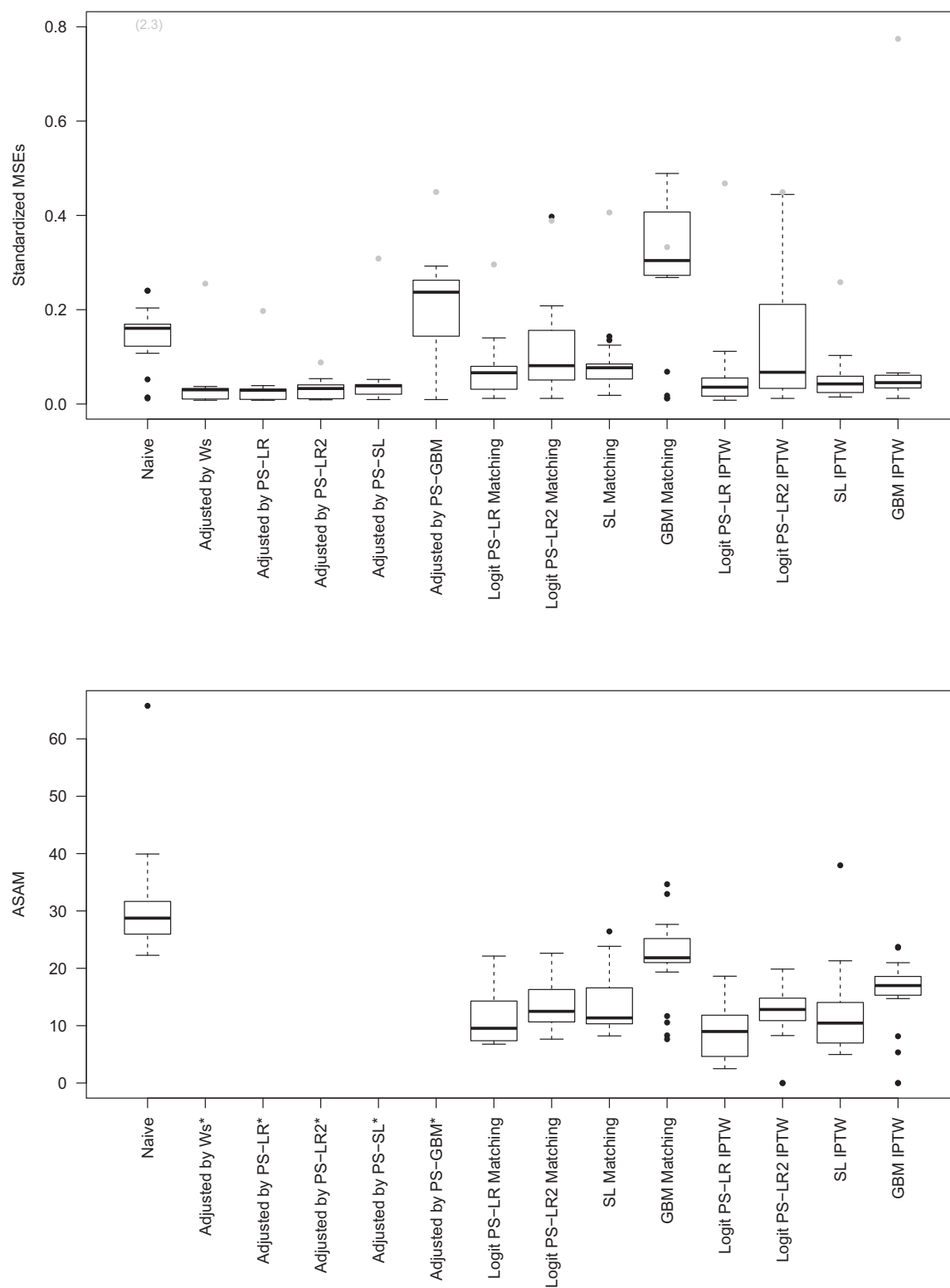


FIGURE 1 Summary of simulations using the synthetic data, aggregated across 21 data generating scenarios. Top: mean squared error (MSE) by analysis method. Gray points and number refer to the MSEs scaled by 1000 for scenario l21. Bottom: average standardized absolute mean difference (ASAM) by analysis method. *Balance statistics for adjusted methods are not possible to compute

In a matched analysis, a greater number of people discarded due to an inability to find a match indicates potential violations of the positivity assumption. The greatest number of people discarded during matching were when matching was being performed using the propensity scores fit by logistic regression with all two-way interactions, and with generalized boosted models, even in larger samples.

The distribution of the weights for the inverse weighted analyses is shown in the Online Supplementary Materials. As the data were sparse, there were occasionally large, and sometimes infinite, weights. In cases of infinite weights, we set the weight to be the maximum of 2 and the largest finite weight in the sample.

The more complex propensity score models, fit via logistic regression with all two-way interactions or generalized boosted models, showed even higher predictive accuracy than the ensemble approach, though all methods were comparable in the largest sample size. This underscores the point that high predictive accuracy is not a desirable feature in a propensity score model. Rather, *balance* of confounding variables across exposure groups is the aim of a propensity score analysis.

4.2 | Results: synthetic data simulations

In the Online Supplementary Materials, we detail the results of each of the synthetic data simulations; here, we aim to give an overview of our findings. We plotted boxplots of standardized MSEs and ASAMs across all 21 scenarios, where the MSEs were standardized by dividing by the square of the true ATEs for each method and for each scenario to be comparable.

The top panel of Figure 1 shows that propensity score regression produces the lowest standardized MSEs with tighter distributions, followed by the inverse weighted approach, with the exception of the case where the propensity score is estimated via a logistic regression with all two-way interactions. Matching provides consistently higher MSEs than the other approaches. Further, within any analytic approach, we see little or no benefit from adopting complex modeling of the propensity score. It is not possible to compute balance statistics for the regression approaches. We observed that, whether matching or inverse weighting, all propensity score estimation methods lead to better balance than the naive (unadjusted) model. However, adding complexity to the fit of the propensity score does not appear to improve covariate balance to any appreciable degree.

5 | DISCUSSION

Researchers in epidemiology and medicine have embraced propensity based methods as a tool for estimating the average effects of an exposure of interest. Considerable research efforts have been devoted to understanding the impact of misspecification of the propensity score model,^{5,6,9,32} and more recently, authors have suggested that the need for less parametric approaches to the widely used approach of logistic regression.^{13,33} Pirracchio et al¹¹ studied the use of ensemble package Super Learner to estimate propensity scores and concluded that it performed better than logistic regression using propensity score matching and weighting.

One of the key findings of this paper is that ensemble methods and other complex models do not perform appreciably better (in terms of MSE) than much simpler approaches such as logistic regression with main effects only. The advantages of using a simpler propensity score model as opposed to a more complex method such as Super Learner or generalized boosting is that the statistical properties of the resulting estimators are much more straightforward to derive either analytically or numerically using bootstrap procedures. It is also the case that simpler models are more interpretable when it comes to understanding the treatment-confounder relationship. Finally, in our experience, more complex propensity score modeling approaches are more prone to producing positivity violations as they focus—in most cases—on prediction of the treatment only.

In our simulations, we used an additive treatment effect model as in (2) and confirmed that, for such a model, direct adjustment is superior to either IPTW or matching, which agrees with other recent literature into propensity score methods.^{20,31} A more general model proposes that the linear predictor takes the form $\mathbf{w}_0\alpha_0 + a\mathbf{w}_1\psi$ for distinct confounder vectors $\mathbf{w}_0$ and $\mathbf{w}_1$. The superiority of direct adjustment holds when the dependence of the outcome on the treatment and treatment-confounder interactions are correctly specified in the proposed outcome model, that is, the difference in expected outcome between treated ($a = 1$) and untreated ($a = 0$) is correctly specified in the conditional model as taking the form $\mathbf{w}_1\psi$. If correct specification does not hold, eg, an interaction is missed in the outcome specification, direct adjustment estimators can be biased. Matching and inverse weighted estimators need not specify interactions for

consistency, and hence offer some robustness in this regard. Of course, doubly robust procedures are preferable still, as these can produce consistent estimators provided either the outcome regression model or the propensity score model is correctly specified, whereas direct regression adjustment approaches, in general, do not.

Our concern in this paper has been deliberately narrow in focussing on the construction of the propensity score for use in a *singly robust* procedure that yields consistent estimation of the ATE if the propensity score model is correctly specified. We note that the use of doubly robust semiparametric procedures that build in additional safeguards by proposing a model for the conditional mean of the outcome given the treatment and confounders can compensate for poor performance of the propensity score estimator. It is equally evident from the literature cited in our Introduction that these advanced methods are currently not widely adopted in practice due to challenges of implementation, whereas advanced methods for modeling the propensity score alone are much more readily implementable using standard software. However, singly robust weighting methods are very widely used in practice. For instance, in a brief survey of all applied research articles published in 2017 in six leading epidemiology journals (American Journal of Epidemiology, Epidemiology, International Journal of Epidemiology, Clinical Infectious Diseases, AIDS, and Lancet HIV) on propensity score estimation for the purpose of IPTW found no instances of doubly robust estimation in the primary analyses (one in a sensitivity analysis), though more than a third employed some form of flexible modeling in the propensity score.

In this paper, we do not address the issue of variable selection in the propensity score model; there is a degree of automated selection in the machine learning and ensemble approaches, but our simpler adjustment approaches rely on fixed specifications. Variable selection is a challenging task in the two-stage (outcome model/propensity) model setting and although the objectives of such selection are theoretically clear—we aim to include in the propensity model only confounders, and in the outcome model pure predictors of outcome—operationalizing these principles is currently the focus of much research. In the context of this paper, however, it is a separate issue, and we do not address it in any detail.

Our simulations are extensive, but not exhaustive, and the results of our simulations may not generalize beyond the settings and candidate methods, and the associated tuning parameters, used in this study (in general, ensemble approaches depend on the choice of the candidate algorithms). A very recent study³⁴ developed an alternative version of an ensemble method that implements a measure of covariate imbalance as the loss function, thus targeting covariate balance rather over predictive accuracy of the propensity score. This approach is appealing, although—like the traditional ensemble approach—carries significant computational burden that may not justify its use over simpler, familiar, and computationally trivial models such as logistic regression. In addition, methods that focus on covariate balance may also enforce balance on covariates that do not have any relationship with the outcome, likely leading to an increase in estimator variance.

ACKNOWLEDGEMENTS

Erica E. M. Moodie and David A. Stephens gratefully acknowledge funding from the Natural Sciences and Engineering Research Council of Canada. Erica E. M. Moodie is also funded by the Fonds de Recherche du Québec - Santé through a Chercheur-Boursier Career award.

ORCID

Erica E. M. Moodie  <https://orcid.org/0000-0002-7225-3977>

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
2. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39(1):33-38.
3. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.
4. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf*. 2004;13(12):855-857.
5. Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4):523-539.
6. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf*. 2008;17(6):546-555.
7. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010;63(8):826-833.

8. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9(4):403-425.
9. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statist Med*. 2010;29(3):337-346.
10. van der Laan MJ, Polley EC, Hubbard AE. Super Learner. *Stat Appl Genet Mol Biol*. 2007;6(1).
11. Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol*. 2015;181(2):108-119.
12. Diaz I, Kelly J. To balance or not to balance. <http://www.unofficialgoogledatascience.com/2016/06/to-balance-or-not-to-balance.html>. 2016.
13. Neugebauer R, Schmittiel JA, van der Laan M. A case study of the impact of data-adaptive versus model-based estimation of the propensity scores on causal inferences from three inverse probability weighting estimators. *Int J Biostat*. 2015;12(1):131-155.
14. Karim ME, Platt RW. Estimating inverse probability weights using Super Learner when weight-model specification is unknown in a marginal structural Cox model context. *Statist Med*. 2017;36(13):2032-2047.
15. Kreif N, Grieve R, Díaz I, Harrison D. Evaluation of the effect of a continuous treatment: a machine learning approach with an application to treatment for traumatic brain injury. *Health Econ*. 2015;24(9):1213-1228.
16. Karim ME, Petkau J, Gustafson P, Tremlett H. On the application of statistical learning approaches to construct inverse probability weights in marginal structural Cox models: hedging against weight-model misspecification. *Commun Stat Simul Comput*. 2016;46(10):7668-7697.
17. Ju C, Combs M, Lendle SD, et al. Propensity score prediction for electronic healthcare databases using Super Learner and High-dimensional Propensity Score Methods. ArXiv e-prints:1703.02236. 2017. <http://adsabs.harvard.edu/abs/2017arXiv170302236J>
18. Gharibzadeh S, Mansournia MA, Foroushani A, et al. Comparing different propensity score estimation methods for estimating the marginal causal effect through standardization to propensity scores. *Commun Stat Simul Comput*. 2018;47(4):964-976.
19. Moodie EEM, Stephens DA. Treatment prediction, balance and propensity score adjustment. *Epidemiology*. 2017;28(5):e51-e53.
20. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006;98(3):253-259.
21. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. 2011;174(11):1213-1222.
22. De Luna X, Waernbaum I, Richardson TS. Covariate selection for the non-parametric estimation of an average treatment effect. *Biometrika*. 2011;98(4):861-875.
23. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219-226.
24. Ridgeway G, McCaffrey D, Morral A, Burgette L, Griffin BA. Toolkit for weighting and analysis of nonequivalent groups: a tutorial for the twang package. R vignette. Sta. Monica, CA: RAND Corporation. 2016.
25. Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. *Econometrica*. 2008;76(6):1537-1557.
26. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.
27. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J Stat Softw*. 2011;42(7).
28. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statist Med*. 2009;28(25):3083-3107.
29. Harrell Jr FE, Dupont MC. The Hmisc package. R package, version 3-0. 2006.
30. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
31. Ertefaie A, Stephens DA. Comparing approaches to causal inference for longitudinal data: inverse probability weighting versus propensity scores. *Int J Biostat*. 2010;6(2).
32. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*. 1993;49(4):1231-1236.
33. Linden A, Yarnold PR. Using classification tree analysis to generate propensity score weights. *Int J Biostat*. 2017;23(4):703-712.
34. Pirracchio R, Carone M. The balance Super Learner: a robust adaptation of the *Super Learner* to improve estimation of the average treatment effect in the treated based on propensity score matching. *Stat Methods Med Res*. 2016;27(8):2504-2518. <https://doi.org/10.1177/0962280216682055>

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: Alam S, Moodie EEM, Stephens DA. Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. *Statistics in Medicine*. 2019;38:1690-1702. <https://doi.org/10.1002/sim.8075>

APPENDIX

IPTW WEIGHT DISTRIBUTION AND PREDICTION ACCURACY MEASURES OF SIMULATED DATA GENERATION SCENARIOS OF THE FIRST STEPS DATABASE

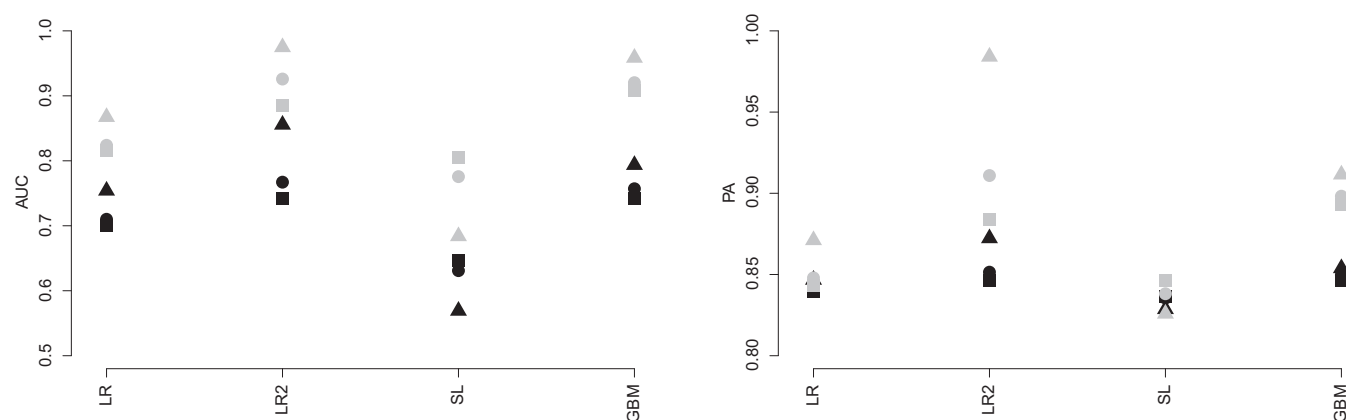


FIGURE A1 Plot showing area under the operating curve (AUC) measure and prediction accuracy (PA) of the scenarios with null treatment effect. Black and gray symbols represent the scenario with four and eight confounders, respectively. Triangles, circles, and squares represent sample size of 100, 300, and 500, respectively. Left: AUC by propensity score (PS) models. Right: PA by PS models

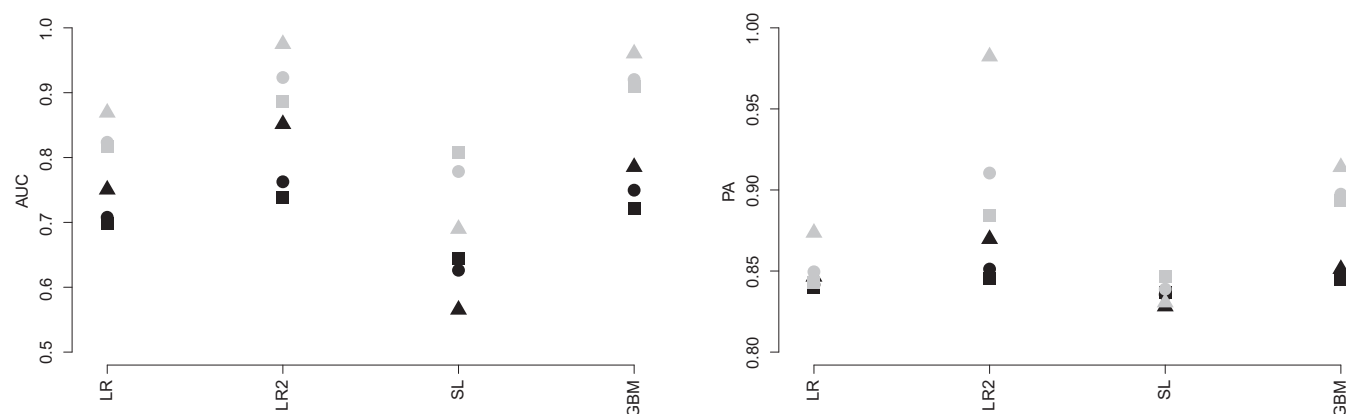


FIGURE A2 Plot showing area under the operating curve (AUC) measure and prediction accuracy (PA) of the scenarios with average treatment effect of 150 g. Black and gray symbols represent the scenario with four and eight confounders, respectively. Triangles, circles, and squares represent sample size of 100, 300, and 500, respectively. Left: AUC by propensity score (PS) models. Right: PA by PS models