Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Should a Propensity Score Model be Super? The Utility of Machine Learning Procedures for Causal Adjustment

Shomoita Alam,[a], Erica E. M. Moodie[a*] and David A. Stephens[b]

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 non-linear settings however experience with real data analyses tend to show overfitting of the propensity score model using this approach. We investigated wider range of 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 and mean squared error, however Super Learner is computationally very expensive and may induce positivity violations in real data settings. Approaches based on generalized boosted models were inferior to both logistic regression and Super Learner in terms of both balance and mean squared error. 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. Finally, we note that to fully understand a complex estimation procedure, simulations based on both real and entirely synthetic data may be needed. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: Average treatment effect, Confounding, Covariate balance, Machine Learning, Propensity Score.

1. Introduction

Randomized control trials are considered to be the gold standard for estimating the causal effect of a treatment (or more generally, an exposure) on an outcome, as their design ensures (in large samples) that there is no confounding of the relationship between treatment and outcome by other covariates, observed or otherwise. Randomization leads to *balance*; essentially, this corresponds to independence of the 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 (non-randomized) 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.

[a] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada*
[b] *Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada*
[*] *Correspondence to: Erica E.M. Moodie, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada. E-mail: erica.moodie@mcgill.ca*

The *propensity score* (Rosenbaum and Rubin, 1983, 1985) can be used to eliminate imbalance in the distribution of the covariates as, within fine 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, or for reweighting samples according to the inverse of their probability of being treated to estimate an average treatment effect. Simple parametric modeling approaches such as logistic regression are widely used to estimate the propensity score (Brookhart et al., 2006), and Rubin (2004) suggested complex models that 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 (Kang and Schafer, 2007).

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, Setoguchi et al. (2008); Westreich et al. (2010); McCaffrey et al. (2004); Lee et al. (2010). Combinations of flexible machine learning methods, termed *ensemble* methods, have also been proposed. Perhaps most prominently, van der Laan et al. (2007) 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 al. (2015) 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 the improvements in estimation were in fact quite modest. Diaz and Kelly (2016) 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. Super Learner has also been considered in the context of longitudinal data, where it has been found useful in the presence of model misspecification (Karim and Platt, 2017). Overall, Super Learner is now being widely adopted in the causal inference literature and in applications (Kreif et al., 2015; Karim et al., 2016; Neugebauer et al., 2016; Ju et al., 2017; Gharibzadeh et al., 2017).

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 (Moodie and Stephens, 2017). Further, the benefits found are often modest, particularly relative to the computational burden of ensemble approaches which 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 nearly two dozen purely synthetic 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 only and using interaction terms, Super Learner, and generalized boosted methods for propensity score modeling in estimating the average treatment effect using propensity score regression, propensity score matching and inverse probability of treatment weighting.

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. The next section examines the results of the real-data simulations in detail, and summarizes 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 that would be classified as X, Z or V if measured can be ignored in the current context.

In the binary treatment context, a propensity score model is a model for the probability of treatment (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
$$g(W) = \Pr[A = 1|W]. \tag{1}$$
This may be achieved by regressing A on W in the observed data.

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,
$$\Pr[A = 1|W, Z] \qquad \Pr[A = 1|W, V]$$
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 treatment. 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 (de Luna et al., 2011). In this paper, we proceed as if all confounders are measured, and 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 its ability to produce balance. Often this is cited as a requirement for correct specification, i.e. 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 re-iterate 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 that are imperfectly specified can achieve adequate adjustment in the sense of providing the desired balance between treatment groups.

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* simulation (Franklin et al., 2014) 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 non-linearities 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 Using Real Data

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 treatment 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 four variables sex, race, parity and smoking status were set to be confounders (W), in that they appeared to be associated with the treatment, 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 eight 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

$$\theta a + \mathbf{w}\alpha \tag{2}$$

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 average treatment effect was set to be either null (0g) or 150g. Variability was added to the outcome in the form of a random mean-zero Normal variable with variance which was the residual variance found in an analysis of the First Steps data (see Online Supplementary Material). Only the 'true' confounders were included. Thus, the four scenarios considered were: (i) four confounders, null ATE; (ii) four confounders, ATE of 150g; (iii) eight confounders, null ATE; or (iv) eight confounders, ATE of 150g.

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 treatment 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 Pirracchio et al. (2015), 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 (2007) and discussed in Diaz and Kelly (2016);
- 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 non-linearly, so that model mis-specification 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 generating 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 Schafer (2007);
- the ensemble approach Super Learner (PS-SL).

The generalized boosted models were implemented using the `twang` package in `R` (Ridgeway et al., 2016). Super Learner was implemented using the same libraries employed by Pirracchio et al. (2015), 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% confidence intervals were computed using a non-parametric bootstrap with 100 resamples. Due to the computational burden of computing bootstrapped standard errors for the ensemble approach, cross-validation was limited to five-fold. The computational time required to calculate the bootstrap standard error from a *single* simulated sample was 20-30 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.

3.4. Estimation of the Average Treatment Effect

The population average treatment effect 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 average treatment effect were considered:

- a *naive model*: regressing the outcome on the treatment only;
- a *covariate-adjusted model*: regressing the outcome on all covariates;
- a *propensity score-adjusted model*: regressing the outcome on the treatment and the propensity score (as a linear, main effect term);
- a *matched analysis*: regressing the outcome on the treatment only, in a sample formed by pair-matching based on the propensity score;
- a *weighted analysis*: regressing the outcome on the treatment only, in a sample weighted by the inverse probability of treatment (Robins et al., 2000).

The matched analysis was carried out using the `Matching` package (Sekhon, 2011), with each subject in the sample is matched once to the nearest subject in the alternative treatment 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 subsection.

The naive model is expected to yield a biased estimator in the presence of confounding, while the covariate adjusted model will provide an unbiased and efficient 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 were compared using the following metrics:

- the bias, empirical standard error, and root mean squared error (rMSE) of the estimated average treatment effect;
- the balance between the covariates between exposed and unexposed subjects (Austin, 2009), 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.

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 (Harrell Jr and Dupont, 2006) and the predictive accuracy, defined as

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

All analyses were performed using R statistical software 3.3.1 (R Core Team, 2016) using high performance computing resources of Compute Canada. R code for the First Steps 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 non-null ATE results are similar, and are provided in the Online Supplementary Materials.

As expected, bias and mean squared error 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 (Ertefaie and Stephens, 2010). 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. The 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.

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 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

Table 1. Simulation Results for the Simulation using First Steps Data with a Null Treatment Effect and Four Confounders, 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(%)[1]	Discarded
Sample Size = 100							
Naive	28.5	170.6	154.1	89.8	173.0	41.5	
Adjusted by Ws	2.3	160.4	160.9	92.4	160.4		
Adjusted by PS-LR	3.2	162.3	164.6	92.6	162.4		
Adjusted by PS-LR2	8.1	179.9	197.7	93.6	180.1		
Adjusted by PS-SL	20.7	163.8	162.1	92.8	165.1		
Adjusted by PS-GBM	11.4	169.8	179.5	95.2	170.2		
Logit PS-LR Matching	10.2	220.2	212.8	95.4	220.4	29.4	10.3
Logit PS-LR2 Matching	10.1	212.5	205.6	93.2	212.7	28.3	31.8
SL Matching	23.1	194.5	203.6	96.4	195.9	52.2	9.1
GBM Matching	13.3	223.3	206.0	94.8	223.7	31.9	15.9
Logit PS-LR IPTW	8.6	196.6	196.3	93.0	196.8	18.0	
Logit PS-LR2 IPTW	16.2	201.6	169.9	89.8	202.3	23.6	
SL IPTW	10.0	302.7	324.9	99.8	302.8	101.0	
GBM IPTW	13.8	187.3	162.6	90.0	187.8	20.3	
Sample Size = 300							
Naive	37.9	94.2	89.7	89.6	101.6	32.4	
Adjusted by Ws	4.8	91.7	91.0	93.2	91.8		
Adjusted by PS-LR	5.3	91.6	91.1	94.0	91.8		
Adjusted by PS-LR2	3.8	95.2	96.5	93.2	95.3		
Adjusted by PS-SL	22.7	91.3	89.9	93.0	94.1		
Adjusted by PS-GBM	1.2	94.7	94.9	92.4	94.7		
Logit PS-LR Matching	4.5	116.0	113.1	93.0	116.1	18.2	4.1
Logit PS-LR2 Matching	2.6	121.0	126.1	95.4	121.0	17.4	15.3
SL Matching	13.6	120.1	120.8	96.2	120.9	27.1	6.7
GBM Matching	3.3	119.5	121.6	95.2	119.6	18.4	7.8
Logit PS-LR IPTW	2.1	104.0	104.0	92.4	104.0	6.7	
Logit PS-LR2 IPTW	0.1	114.1	119.8	94.2	114.1	8.7	
SL IPTW	16.5	105.8	139.9	96.2	107.1	15.4	
GBM IPTW	2.0	105.4	99.4	91.4	105.4	5.0	
Sample Size = 500							
Naive	35.4	70.3	68.6	91.4	78.7	30.5	
Adjusted by Ws	1.7	69.1	69.7	92.4	69.2		
Adjusted by PS-LR	1.9	68.8	69.8	93.4	68.8		
Adjusted by PS-LR2	2.0	70.6	72.2	93.4	70.7		
Adjusted by PS-SL	14.6	68.3	69.1	94.0	69.9		
Adjusted by PS-GBM	0.4	70.8	72.1	93.2	70.8		
Logit PS-LR Matching	1.3	82.3	84.5	94.8	82.3	14.0	3.6
Logit PS-LR2 Matching	2.6	85.9	89.4	94.8	85.9	14.2	11.5
SL Matching	8.0	86.3	92.0	96.4	86.6	21.3	5.2
GBM Matching	0.3	85.6	89.1	94.2	85.6	14.6	6.9
Logit PS-LR IPTW	0.7	74.5	77.1	92.8	74.5	4.7	
Logit PS-LR2 IPTW	2.0	82.3	88.6	94.6	82.4	4.8	
SL IPTW	10.1	76.4	89.5	95.2	77.1	10.8	
GBM IPTW	1.9	77.3	77.5	92.6	77.3	3.1	

[1] Balance for covariates child's sex, mother's race, parity and smoking status.

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 are 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. As the highly parameterized logistic regression and the

Table 2. Simulation Results for the Simulation using First Steps Data with a Null Treatment Effect and Eight Confounders, 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(%)[1]	Discarded
Sample Size = 100							
Naive	46.4	186.9	158.9	87.6	192.6	56.3	
Adjusted by Ws	0.7	178.7	175.5	92.8	178.7		
Adjusted by PS-LR	2.9	181.8	197.4	95.4	181.8		
Adjusted by PS-SL	28.8	177.4	197.1	95.4	179.7		
Adjusted by PS-GBM	38.9	255.5	398.8	95.6	258.5		
Logit PS-LR Matching	4.8	293.6	318.2	98.2	293.6	35.6	21.1
SL Matching	22.5	238.7	299.6	99.0	239.7	54.3	11.3
GBM Matching	29.6	339.8	414.9	99.8	341.1	53.4	64.8
Logit PS-LR IPTW	4.1	270.0	254.2	93.8	270.0	32.1	
SL IPTW	27.7	339.6	354.6	99.6	340.7	71.5	
GBM IPTW	22.0	190.7	150.9	85.2	191.9	37.4	
Sample Size = 300							
Naive	61.2	99.3	91.1	86.4	116.6	49.6	
Adjusted by Ws	3.2	97.2	94.9	91.4	97.2		
Adjusted by PS-LR	4.6	98.5	95.8	90.8	98.6		
Adjusted by PS-LR2	0.4	125.9	166.8	99.4	125.9		
Adjusted by PS-SL	21.2	97.4	105.3	94.4	99.7		
Adjusted by PS-GBM	28.7	126.4	131.4	91.2	129.6		
Logit PS-LR Matching	13.1	176.2	165.8	97.4	176.7	18.9	5.2
Logit PS-LR2 Matching	4.1	271.4	276.8	99.4	271.5	36.5	19.6
SL Matching	7.5	157.6	223.5	99.2	157.8	26.1	4.8
GBM Matching	18.7	252.3	245.6	98.0	253.0	31.9	33.6
Logit PS-LR IPTW	5.0	142.6	135.0	90.4	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	99.8	188.8	23.8	
GBM IPTW	18.8	124.9	99.3	87.8	126.3	20.6	
Sample Size = 500							
Naive	57.7	74.5	69.4	82.6	94.2	48.3	
Adjusted by Ws	3.1	73.9	72.6	92.0	74.0		
Adjusted by PS-LR	5.8	73.4	73.0	92.6	73.6		
Adjusted by PS-LR2	3.1	82.4	95.0	95.6	82.5		
Adjusted by PS-SL	11.2	76.9	83.2	92.6	77.7		
Adjusted by PS-GBM	24.7	93.0	93.3	90.8	96.2		
Logit PS-LR Matching	19.8	128.3	126.4	96.6	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	99.0	130.6	19.2	2.6
GBM Matching	11.8	197.2	196.7	97.6	197.6	26.4	18.4
Logit PS-LR IPTW	12.2	103.3	101.0	92.2	104.1	9.9	
Logit PS-LR2 IPTW	1.0	211.0	257.0	99.6	211.0	27.1	
SL IPTW	13.5	186.3	286.2	99.4	186.8	22.3	
GBM IPTW	21.8	98.8	83.3	90.2	101.1	15.9	

[1] Balance for covariates child's sex, mother's age, race, parity, marital status, smoking status, weight prior to pregnancy and education level.

generalized boosted models did not provide the better estimates of the ATE, this underscores the point that high predictive accuracy is not a desirable feature in a propensity score model. Rather, *balance* of confounding variables across treatment 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 mean squared errors (MSEs) and ASAMs across all 21 scenarios, where the MSEs were standardized by dividing by the square of the true ATE.

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: Super Learner and main-effects logistic regression tend to perform similarly, while generalized boosted models generally perform much worse. 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. Note that it is not possible to compute balance statistics for the regression approaches.

5. Discussion

Researchers in epidemiology and medicine have embraced propensity based methods as a tool for estimating the average effects of an treatment of interest. Considerable research efforts have been devoted to understanding the impact of misspecification of the propensity score model (Drake, 1993; Kang and Schafer, 2007; Setoguchi et al., 2008; Lee et al., 2010), and more recently authors have suggested the need for less parametric approaches to the widely used logistic regression. Pirracchio et al. (2015) studied the use of ensemble package Super Learner to estimate propensity scores and concluded that it performed better than logistic regression when using propensity score matching and IPTW. One of the key findings of the this paper is that ensemble methods and other complex models do not perform appreciably better (in terms of mean squared error) 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.

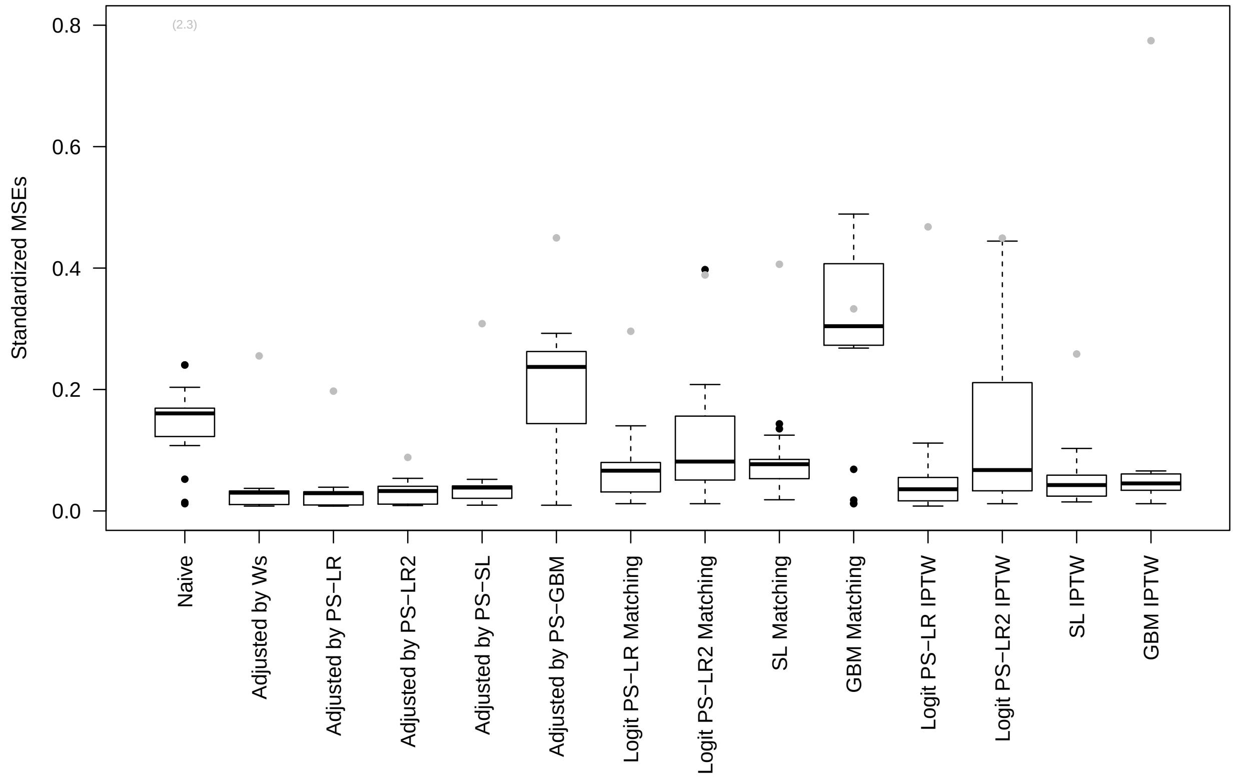
In our simulations, we used an additive treatment effect model as in (2), and confirmed that for such a model direct adjustment is superior to either IPTW or matching, which agrees with other recent literature into propensity score methods (Ertefaie and Stephens, 2010; Moodie and Stephens, 2017). A more 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, then doubly robust inverse weighting procedures can still 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. However, the results we have obtained concerning the propensity score model and its construction still hold in this more complicated setting.

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 study (Pirracchio and Carone, 2016) 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 positivity violations and estimator variance.



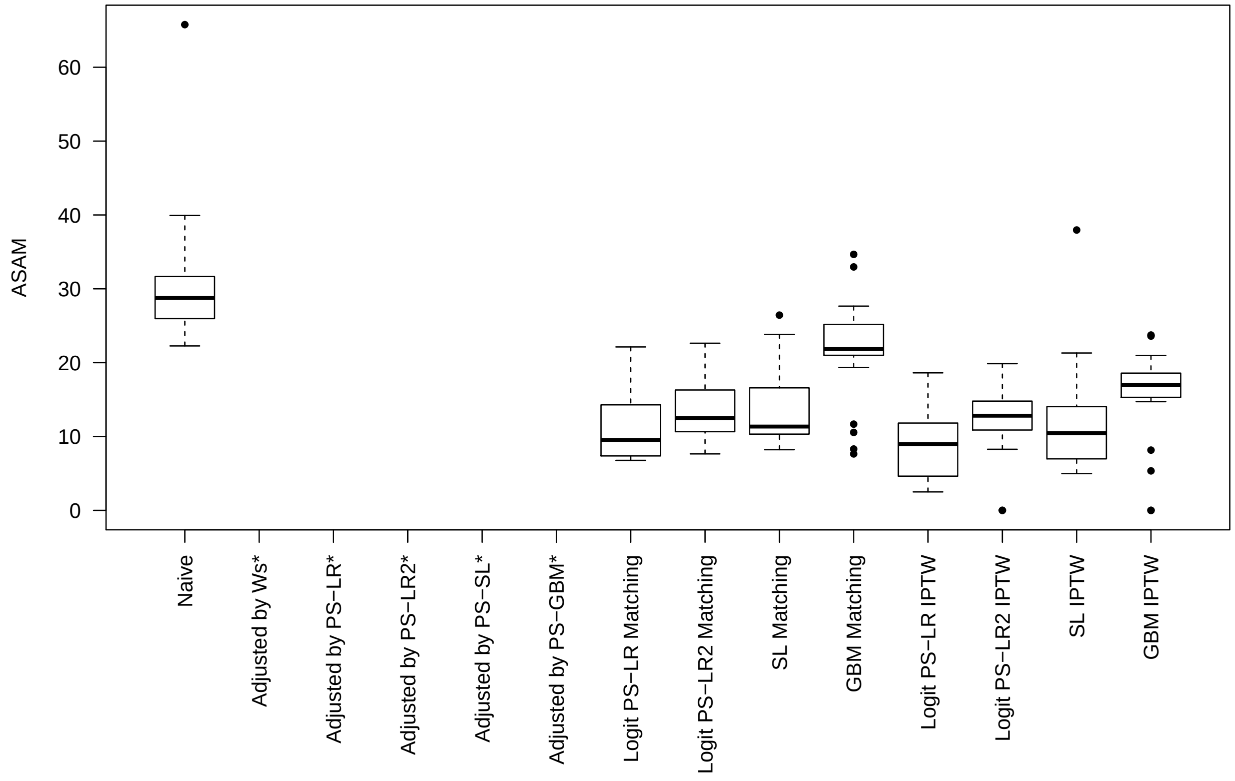


Figure 1. Summary of Simulations Using the Synthetic Data, Aggregated Across 21 Data Generating Scenarios. Top: Mean Squared Error by Analysis Method. Grey Points and Number Refer to the MSEs Scaled by 1000 for Scenario S21. Bottom: Average Standardized Absolute Mean Difference (ASAM) by Analysis Method. Note: Balance Statistics are not Available for Regression Adjusted Models.

Appendix. Measures of Prediction Accuracy in First Steps Simulation Study

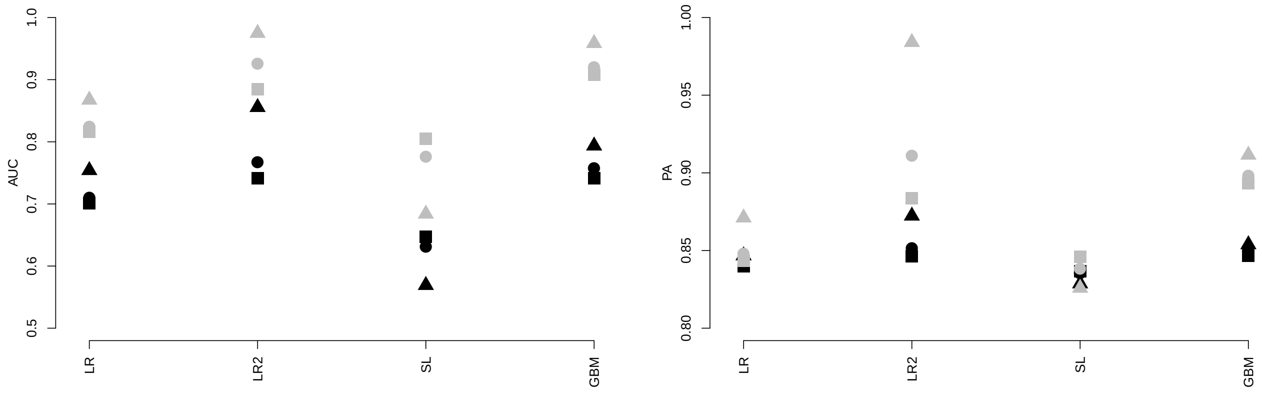


Figure 2. Plot Showing Area Under the Operating Curve (AUC, left column) and Prediction Accuracy (PA, right column) of the Null Treatment Effect Scenarios. Black and Grey Symbols Represent the Scenario with Four and Eight Confounders, Respectively. Triangles, Circles, and Squares Represent Sample Size of 100, 300, and 500, Respectively.

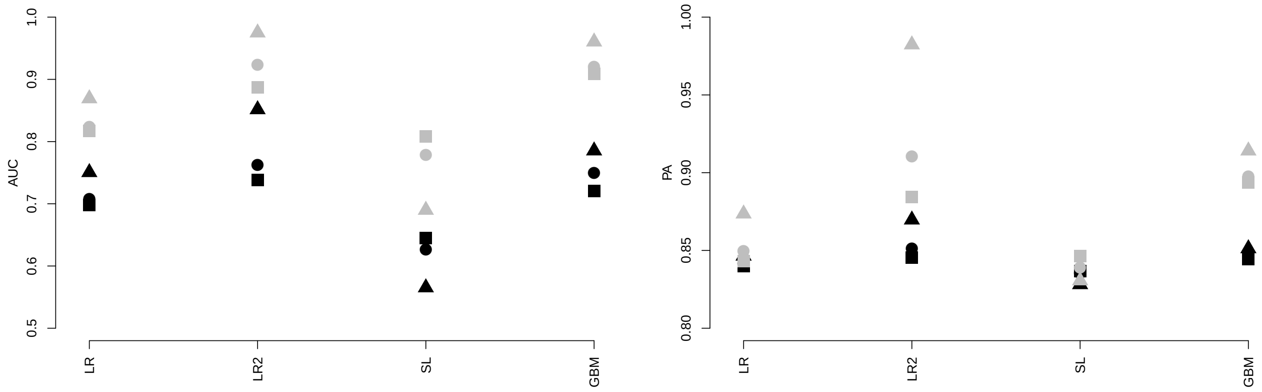


Figure 3. Plot Showing Area Under the Operating Curve (AUC, left column) and Prediction Accuracy (PA, right column) of the ATE=150g Scenarios. Black and Grey Symbols Represent the Scenario with Four and Eight Confounders, Respectively. Triangles, Circles, and Squares Represent Sample Size of 100, 300, and 500, Respectively.

References

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156.

de Luna, X., Waernbaum, I., and Richardson, T. (2011). Covariate selection for the non-parametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Diaz, I. and Kelly, J. (2016). To balance or not to balance? http://www.unofficialgoogledatascience.com/2016/06/to-balance-or-not-to-balance.html.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49(4):1231–1236.

Ertefaie, A. and Stephens, D. A. (2010). Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *The International Journal of Biostatistics*, 6(2).

Franklin, J. M., Schneeweiss, S., Polinski, J. M., and Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis*, 72:219–226.

Gharibzadeh, S., Mansournia, M. A., Foroushani, A., Alizadeh, A., Amouzegar, A., Mehrabani-Zeinabad, K., and Mohammad, K. (2017). Comparing different propensity score estimation methods for estimating the marginal causal effect through standardization to propensity scores. *Communications in Statistics – Simulation and Computation*, (in press).

Harrell Jr, F. E. and Dupont, M. C. (2006). The Hmisc package. *R package version*, pages 3–0.

Ju, C., Combs, M., Lendle, S. D., Franklin, J. M., Wyss, R., Schneeweiss, S., and van der Laan, M. J. (2017). Propensity score prediction for electronic healthcare dataset using super learner and high-dimensional propensity score method. (arXiv:1703.02236v2 [stat.AP]).

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

Karim, M. E., Petkau, J., Gustafson, P., and Tremlett, H. (2016). On the application of statistical learning approaches to construct inverse probability weights in marginal structural Cox models: Hedging against weight-model misspecification. *Communications in Statistics – Simulation and Computation*, (in press).

Karim, M. E. and Platt, R. W. (2017). Estimating inverse probability weights using Super Learner when weight-model specification is unknown in a marginal structural Cox model context. *Statistics in Medicine (in press)*.

Kreif, N., Grieve, R., Díaz, I., and Harrison, D. (2015). Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury. *Health economics*, 24(9):1213–1228.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403.

Moodie, E. E. and Stephens, D. A. (2017). Treatment prediction, balance and propensity score adjustment. *Epidemiology*, 28(5):e51–53.

Neugebauer, R., Schmittdiel, J. A., and van der Laan, M. J. (2016). 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. *The International Journal of Biostatistics*, 12(1):131–155.

Pirracchio, R. and Carone, M. (2016). 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. *Statistical Methods in Medical Research*, page doi: 10.1177/0962280216682055.

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

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., and Griffin, B. A. (2016). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. *R vignette. RAND.*

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, 13(12):855–857.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *Journal of Statistical Software*, 42(7).

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, 17(6):546–555.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833.

Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

SUPPLEMENTARY MATERIALS FOR "Should a Propensity Score Model be Super? – The Utility of Ensemble Procedures for Causal Adjustment"

Shomoita Alam,[a], Erica E. M. Moodie[a*] and David A. Stephens[b]

1. Simulation for Fully Synthetic Data Analysis

1.1. Data Generation and Estimation: Scenarios S1-S16

We explored 21 synthetic data simulation scenarios with models of varying degrees of complexity in the propensity score. We shall refer to these scenarios are S1 - S21. More than two thirds of the scenarios (S1-S16) follow the approach of Pirracchio et al. (2015), who follow the work of Setoguchi et al. (2008).

Data Generation For each of the first several settings, we have the following covariates: $\mathbf{W}$ is a vector of 4 confounders: $(\mathbf{W_1} - \mathbf{W_4})$, $\mathbf{Z}$ is the vector of 3 instruments: $(\mathbf{Z_5} - \mathbf{Z_7})$, and $\mathbf{X}$ is the vector of 3 outcome predictors or risk factors $(\mathbf{X_8} - \mathbf{X_{10}})$. These ten covariates are generated as follows:

(a) First, 8 standard normal random variables $(\mathbf{V_i},\ i = 1, \ldots, 6, 8, 9)$ were generated.

(b) Then, 8 covariates $(\mathbf{W_i},\ i = 1, \ldots, 4;\ \mathbf{Z_i},\ i = 5, 6;\ \mathbf{X_i},\ i = 8, 9)$ are calculated as a linear combination of $\mathbf{V_i}$, $i = 1, \ldots, 6, 8, 9$, with correlation ranging from from 0.2 to 0.9 introduced between some of the variables.

(c) A further 2 covariates $(\mathbf{Z_7}, \mathbf{X_{10}})$ were generated as independent standard normal random variables.

(d) Finally, 6 of the 10 covariates $(\mathbf{W_1}, \mathbf{W_3}, \mathbf{Z_5}, \mathbf{Z_6}, \mathbf{X_8}, \mathbf{X_9})$ were dichotomized based on the mean value of each covariate.

Next, the treatment variable was generated. The probability that the exposure A equaled 1 was generated as a function of the covariates $\mathbf{W_i}, i = 1, \ldots, 4$ and $\mathbf{Z_i}, i = 5, \ldots, 7$ (i.e. only the confounders and instruments), according to

$$\Pr[A = 1|\mathbf{W_i}, \mathbf{Z_i}] = (1 + \exp(-(\beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 Z_5 + \beta_6 Z_6 + \beta_7 Z_7)))^{-1},$$

for S1-S4, and according to a non-linear model:

$$\begin{aligned}\Pr[A = 1|\mathbf{W_i}, \mathbf{Z_i}] = (1 + \exp(-(&\beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 Z_5 + \beta_6 Z_6 + \beta_7 Z_7 \\ &+ \beta_2 W_2 W_2 + \beta_4 W_4 W_4 + \beta_7 Z_7 Z_7)))^{-1}\end{aligned}$$

[a] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada*
[b] *Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada*
[*] *Correspondence to: Shomoita Alam, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada. E-mail: shomoita.alam@mail.mcgill.ca*

Table 1. Coefficients for the treatment model for Scenario S1-S16

	Value for Scenarios	
Coefficient	S1-S2, S5-S6 S9-S10, S13-S14	S3-S4, S7-S8 S11-S12, S15-S16
β_0	0.00	0.00
β_1	0.80	0.80
β_2	-0.25	-0.25
β_3	0.60	0.60
β_4	-0.40	-0.40
β_5	-0.80	-1.60
β_6	-0.50	-0.50
β_7	0.70	1.40

for S5-S8. In Scenarios S9-S12, non-additivity was considered in the form of two-way interactions:

$$\begin{aligned}\Pr[A=1|\mathbf{W_i},\mathbf{Z_i}] = (1+\exp(-(&\beta_0+\beta_1W_1+\beta_2W_2+\beta_3W_3+\beta_4W_4+\beta_5Z_5+\beta_6Z_6+\beta_7Z_7\\ &+0.5\beta_1W_1W_3+0.7\beta_2W_2W_4+0.5\beta_3W_3Z_5+0.7\beta_4W_4Z_6+0.5\beta_5Z_5Z_7+0.5\beta_1W_1Z_6\\ &+0.7\beta_2W_2W_3+0.5\beta_3W_3W_4+0.5\beta_4W_4Z_5+0.5\beta_5Z_5Z_6)))^{-1}.\end{aligned}$$

Non-additivity and non-linearity together were considered in S13-S16:

$$\begin{aligned}\Pr[A=1|\mathbf{W_i},\mathbf{Z_i}] = (1+\exp(-(&\beta_0+\beta_1W_1+\beta_2W_2+\beta_3W_3+\beta_4W_4+\beta_5Z_5+\beta_6Z_6+\beta_7Z_7\\ &+\beta_2W_2W_2+\beta_4W_4W_4+\beta_7Z_7Z_7\\ &+0.5\beta_1W_1W_3+0.7\beta_2W_2W_4+0.5\beta_3W_3Z_5+0.7\beta_4W_4Z_6+0.5\beta_5Z_5Z_7+0.5\beta_1W_1Z_6\\ &+0.7\beta_2W_2W_3+0.5\beta_3W_3W_4+0.5\beta_4W_4Z_5+0.5\beta_5Z_5Z_6)))^{-1}.\end{aligned}$$

The values of β are given in Table 1.

The continuous outcome variable Y as a linear function of A and $\mathbf{W_i}$ and $\mathbf{X_i}$ (only the confounders and outcome predictors):

$$Y=\alpha_0+\alpha_i\mathbf{W_i}+\alpha_i\mathbf{X_i}+\gamma A+\epsilon$$

where the effect of exposure, $\gamma=-0.4$, and $\epsilon\sim N(0,0.09)$. The α coefficients of the outcome model are given by (-3.85,0.3,-0.36,-0.73,-0.2,0.71,-0.19,0.26).

Analytic models For the estimation of the propensity scores in odd-numbered settings (Scenarios S1, S3, ..., S15), we used *only* the covariates which are confounders and outcome predictors, and excluded instruments from the models. In contrast, for even-numbered settings (Scenarios S2, S4, ..., S16), all covariates (confounders, instruments, and outcome predictors) were included in the treatment model.

For the estimation of the outcome models which is only adjusted by the covariates, we used all the of them (confounders, instruments, and outcome predictors).

1.2. Data Generation and Estimation: Scenario S17

In this setting, we considered only two binary covariates, W_1 and W_2, each drawn from a Bernoulli(0.6) distribution. The treatment model was given by:

$$\Pr[A=1|W_1,W_2]=\begin{cases}0.02 & \text{if } W_1=W_2=0\\ 0.48 & \text{if } W_1=0,W_2=1\\ 0.20 & \text{if } W_1=1,W_2=0\\ 0.30 & \text{if } W_1=W_2=1\end{cases}.$$

Here, W_1 and W_2 are confounders and Y can depend on W_1 and W_2 as main effects only. But the product/interaction of W_1 and W_2 is not a confounder and appears in the model for treatment but not outcome. The coefficients for the outcome model are given by an intercept of -3.85, and 0.3 and -0.36 for W_1 and W_2, respectively. Both covariates were used in the analytic model to estimate the propensity score.

1.3. Data Generation and Estimation: Scenario S18

In this setting, we let $\mathbf{W}$ be a vector of 10 binary covariates each drawn independently from a $Bernoulli(0.45)$ distribution. The treatment is drawn from a Bernoulli distribution whose probability distribution depends on all the main effects of the 10 covariates and two 4-way interactions ($W1 : W2 : W3 : W4$) and ($W3 : W4 : W5 : W6$) with their lower order interactions included in the model. As in Scenario 17, the interaction terms are intended to act like instruments in the model. The coefficients of the treatment and outcome models are provided in Table 2. All covariates were used in the analytic model to estimate the propensity score.

Table 2. Coefficients for the treatment and outcome models for Scenario S18

Treatment model		Outcome model	
Coefficient	Value	Coefficient	Value
β_0	0	α_0	-3.85
β_1	0.8	α_1	0.3
β_2	-0.25	α_2	-0.36
β_3	0.6	α_3	-0.73
β_4	-0.4	α_4	-0.2
β_5	-0.8	α_5	0.71
β_6	-0.5	α_6	-0.19
β_7	0.7	α_7	0.26
β_8	0.3	α_8	0.7
β_9	-0.01	α_9	-0.09
β_{10}	0.1	α_{10}	0.4
β_{1234}	0.9		
β_{123}	0.01		
β_{124}	0.02		
β_{134}	0.03		
β_{234}	0.04		
β_{12}	0.001		
β_{13}	0.002		
β_{14}	0.003		
β_{23}	0.004		
β_{24}	0.005		
β_{34}	0.006		
β_{3456}	-1.2		
β_{345}	-0.01		
β_{346}	0.02		
β_{356}	-0.03		
β_{456}	0.04		
β_{34}	-0.001		
β_{35}	0.002		
β_{36}	-0.003		
β_{45}	-0.004		
β_{46}	-0.005		
β_{56}	0.006		

1.4. Data Generation and Estimation: Scenarios S19-S20

In this setting, $\mathbf{W}$ is the vector of two $Bernoulli(0.5)$ random variables. The treatment is also a Bernoulli random variable with probability

$$\Pr[A = 1|\mathbf{W_i}] = \frac{\exp(\beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_1 W_2)}{1 + \exp(\beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_1 W_2)},$$

where coefficients β are (-2, 3, -2.5, 2.5) in S19 and (-0.25, 1, 1, -4) in S20. The coefficients for the outcome model are as in S18: an intercept of -3.85, and 0.3 and -0.36 for W_1 and W_2, respectively. Both covariates were used in the analytic model to estimate the propensity score.

1.5. *Data Generation and Estimation: Scenario S21*

Finally, we adopted a setting from the simulation study conducted in Kang and Schafer (2007) and Diaz and Kelly (2016). The true set of confounders, (U_1, U_2, U_3, U_4), is generated independently and identically distributed, from a Normal distribution with mean 0 and a diagonal covariance matrix with variances of 1. Further, six instrumental variables $(Z_5 - Z_{10})$ are independently generated from a $Bernoulli(0.5)$. The treatment is a Bernoulli random variable with probability

$$\begin{aligned}\Pr[A = 1|\mathbf{U_i}, \mathbf{Z_i}] = \text{expit}(& -U_1 + 0.5U_2 - 0.25U_3 - 0.1U_4 + .1Z_5 + 0.3Z_6 - 0.7Z_7 + 1.2Z_8 \\ & + .2Z_9 - .3Z_{10}).\end{aligned}$$

The continuous outcome variable Y was generated from a linear combination of A and $\mathbf{U_i}$ (main effects only):

$$Y = 210 + 27.4U_1 + 13.7U_2 + 13.7U_3 + 13.7U_4 + \gamma A + \epsilon$$

where the effect of exposure, $\gamma = -0.4$, and $\epsilon \sim N(0, 1)$.

Again following Kang and Schafer (2007), we transformed the true confounders to create variables $W_1 - W_4$ as follows:

$$\begin{aligned} W_1 &= \exp(U_1/2), \\ W_2 &= U_2/(1 + \exp(U_1)), \\ W_3 &= (((U_1U_3)/25) + 0.6)^3, \\ W_4 &= (U_2 + U_4 + 20)^2. \end{aligned}$$

These variables, rather than $\mathbf{U_i}$, were made available for the analysis, so that in the estimation, the (mis-specified) propensity score was fit as $A \sim W_1 + W_2 + W_3 + W_4 + Z_5 + Z_6 + Z_7 + Z_8 + Z_9 + Z_{10}$.

Performance was assessed using 500 replicated data sets of size $n = 300$ for all simulation scenarios (S1-S21).

1.6. *Detailed Results for the Simulation Scenarios of Synthetic Data*

In this section, we provide the full results for the synthetic data simulations, which are summarized in Figure 1 of the main text.

Table 3. Simulation Results for Scenario S1

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.225	0.175	0.089	0.196	28.749	
Adjusted by Ws	-0.399	0.001	0.039	0.039		
Adjusted by PS-LR	-0.398	0.002	0.037	0.038		
Adjusted by PS-LR2	-0.399	0.001	0.038	0.038		
Adjusted by PS-SL	-0.345	0.055	0.053	0.076		
Adjusted by PS-GBM	-0.592	0.192	0.091	0.212		
Logit PS-LR Matching	-0.396	0.004	0.066	0.066	15.630	4.944
Logit PS-LR2 Matching	-0.401	0.001	0.076	0.076	15.931	8.628
SL Matching	-0.340	0.060	0.080	0.100	19.325	2.738
GBM Matching	-0.566	0.166	0.132	0.212	27.632	45.920
Logit PS-LR IPTW	-0.399	0.001	0.043	0.043	12.429	
Logit PS-LR2 IPTW	-0.396	0.004	0.054	0.054	13.769	
SL IPTW	-0.395	0.005	0.054	0.054	14.193	
GBM IPTW	-0.342	0.058	0.054	0.079	18.583	

Table 4. Simulation Results for Scenario S2

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.225	0.175	0.089	0.196	28.749	
Adjusted by Ws	-0.399	0.001	0.039	0.039		
Adjusted by PS-LR	-0.400	0.000	0.039	0.039		
Adjusted by PS-LR2	-0.401	0.001	0.042	0.042		
Adjusted by PS-SL	-0.366	0.034	0.047	0.058		
Adjusted by PS-GBM	-0.572	0.172	0.091	0.195		
Logit PS-LR Matching	-0.402	0.002	0.082	0.082	7.324	7.486
Logit PS-LR2 Matching	-0.390	0.010	0.110	0.110	10.514	10.820
SL Matching	-0.369	0.031	0.077	0.083	9.946	3.416
GBM Matching	-0.546	0.146	0.147	0.207	20.997	61.634
Logit PS-LR IPTW	-0.397	0.003	0.061	0.061	4.632	
Logit PS-LR2 IPTW	-0.389	0.011	0.122	0.123	10.794	
SL IPTW	-0.395	0.005	0.064	0.065	5.841	
GBM IPTW	-0.326	0.074	0.061	0.096	15.773	

Table 5. Simulation Results for Scenario S3

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.264	0.136	0.091	0.163	31.661	
Adjusted by Ws	-0.400	0.000	0.045	0.045		
Adjusted by PS-LR	-0.398	0.002	0.038	0.038		
Adjusted by PS-LR2	-0.398	0.002	0.039	0.040		
Adjusted by PS-SL	-0.332	0.068	0.061	0.091		
Adjusted by PS-GBM	-0.579	0.179	0.098	0.204		
Logit PS-LR Matching	-0.400	0.000	0.062	0.062	22.129	4.364
Logit PS-LR2 Matching	-0.393	0.007	0.070	0.071	22.632	8.854
SL Matching	-0.331	0.069	0.081	0.106	26.438	3.042
GBM Matching	-0.562	0.162	0.137	0.212	34.664	48.906
Logit PS-LR IPTW	-0.398	0.002	0.041	0.041	18.619	
Logit PS-LR2 IPTW	-0.396	0.004	0.051	0.051	19.864	
SL IPTW	-0.382	0.018	0.053	0.056	21.308	
GBM IPTW	-0.358	0.042	0.056	0.070	23.599	

Table 6. Simulation Results for Scenario S4

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.264	0.136	0.091	0.163	31.661	
Adjusted by Ws	-0.400	0.000	0.045	0.045		
Adjusted by PS-LR	-0.401	0.001	0.046	0.046		
Adjusted by PS-LR2	-0.401	0.001	0.052	0.052		
Adjusted by PS-SL	-0.379	0.021	0.053	0.057		
Adjusted by PS-GBM	-0.487	0.087	0.090	0.125		
Logit PS-LR Matching	-0.390	0.010	0.120	0.120	11.062	8.348
Logit PS-LR2 Matching	-0.396	0.004	0.182	0.183	17.855	10.784
SL Matching	-0.375	0.025	0.112	0.115	12.114	3.994
GBM Matching	-0.473	0.073	0.198	0.211	22.035	75.386
Logit PS-LR IPTW	-0.399	0.001	0.109	0.109	10.097	
Logit PS-LR2 IPTW	-0.386	0.014	0.192	0.193	17.502	
SL IPTW	-0.393	0.007	0.099	0.099	10.446	
GBM IPTW	-0.330	0.070	0.069	0.099	19.606	

Table 7. Simulation Results for Scenario S5

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.285	0.115	0.118	0.165	24.709	
Adjusted by Ws	-0.401	0.001	0.072	0.072		
Adjusted by PS-LR	-0.406	0.006	0.068	0.069		
Adjusted by PS-LR2	-0.407	0.007	0.072	0.072		
Adjusted by PS-SL	-0.380	0.020	0.078	0.080		
Adjusted by PS-GBM	-0.565	0.165	0.123	0.206		
Logit PS-LR Matching	-0.404	0.004	0.094	0.094	13.060	4.678
Logit PS-LR2 Matching	-0.406	0.006	0.113	0.113	13.971	9.330
SL Matching	-0.380	0.020	0.107	0.109	16.191	3.462
GBM Matching	-0.542	0.142	0.167	0.220	23.349	45.138
Logit PS-LR IPTW	-0.411	0.011	0.071	0.072	9.796	
Logit PS-LR2 IPTW	-0.406	0.006	0.089	0.089	11.838	
SL IPTW	-0.416	0.016	0.079	0.081	11.904	
GBM IPTW	-0.376	0.024	0.081	0.084	15.925	

Table 8. Simulation Results for Scenario S6

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.285	0.115	0.118	0.165	24.709	
Adjusted by Ws	-0.401	0.001	0.072	0.072		
Adjusted by PS-LR	-0.402	0.002	0.073	0.073		
Adjusted by PS-LR2	-0.404	0.004	0.083	0.083		
Adjusted by PS-SL	-0.405	0.005	0.080	0.080		
Adjusted by PS-GBM	-0.552	0.152	0.129	0.199		
Logit PS-LR Matching	-0.408	0.008	0.112	0.112	6.773	5.776
Logit PS-LR2 Matching	-0.393	0.007	0.150	0.151	9.533	10.578
SL Matching	-0.405	0.005	0.115	0.115	8.212	8.722
GBM Matching	-0.539	0.139	0.209	0.251	19.348	75.172
Logit PS-LR IPTW	-0.403	0.003	0.083	0.083	3.360	
Logit PS-LR2 IPTW	-0.391	0.009	0.174	0.175	10.879	
SL IPTW	-0.405	0.005	0.095	0.095	5.577	
GBM IPTW	-0.360	0.040	0.090	0.099	14.716	

Table 9. Simulation Results for Scenario S7

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.317	0.083	0.112	0.140	25.295	
Adjusted by Ws	-0.400	0.000	0.070	0.070		
Adjusted by PS-LR	-0.408	0.008	0.065	0.065		
Adjusted by PS-LR2	-0.410	0.010	0.068	0.069		
Adjusted by PS-SL	-0.373	0.027	0.077	0.082		
Adjusted by PS-GBM	-0.549	0.149	0.132	0.199		
Logit PS-LR Matching	-0.408	0.008	0.095	0.096	16.659	3.994
Logit PS-LR2 Matching	-0.409	0.009	0.104	0.105	17.456	9.350
SL Matching	-0.375	0.025	0.107	0.110	20.629	3.276
GBM Matching	-0.533	0.133	0.181	0.224	27.657	48.144
Logit PS-LR IPTW	-0.411	0.011	0.066	0.067	13.019	
Logit PS-LR2 IPTW	-0.414	0.014	0.087	0.088	14.795	
SL IPTW	-0.409	0.009	0.074	0.074	15.782	
GBM IPTW	-0.388	0.012	0.077	0.078	18.173	

Table 10. Simulation Results for Scenario S8

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.317	0.083	0.112	0.140	25.295	
Adjusted by Ws	-0.400	0.000	0.070	0.070		
Adjusted by PS-LR	-0.401	0.001	0.071	0.071		
Adjusted by PS-LR2	-0.404	0.004	0.081	0.081		
Adjusted by PS-SL	-0.411	0.011	0.083	0.084		
Adjusted by PS-GBM	-0.500	0.100	0.142	0.174		
Logit PS-LR Matching	-0.410	0.010	0.121	0.122	7.372	7.942
Logit PS-LR2 Matching	-0.388	0.012	0.160	0.160	10.653	10.016
SL Matching	-0.413	0.013	0.151	0.152	10.324	14.480
GBM Matching	-0.488	0.088	0.256	0.271	21.344	91.540
Logit PS-LR IPTW	-0.407	0.007	0.104	0.105	6.209	
Logit PS-LR2 IPTW	-0.406	0.006	0.243	0.243	16.356	
SL IPTW	-0.394	0.006	0.105	0.105	8.523	
GBM IPTW	-0.364	0.036	0.094	0.100	17.017	

Table 11. Simulation Results for Scenario S9

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.262	0.138	0.117	0.181	30.981	
Adjusted by Ws	-0.398	0.002	0.073	0.073		
Adjusted by PS-LR	-0.404	0.004	0.071	0.071		
Adjusted by PS-LR2	-0.407	0.007	0.075	0.075		
Adjusted by PS-SL	-0.378	0.022	0.076	0.079		
Adjusted by PS-GBM	-0.554	0.154	0.123	0.197		
Logit PS-LR Matching	-0.415	0.015	0.108	0.109	14.290	5.204
Logit PS-LR2 Matching	-0.408	0.008	0.122	0.122	15.095	8.946
SL Matching	-0.379	0.021	0.111	0.113	16.587	2.962
GBM Matching	-0.530	0.130	0.171	0.215	25.183	44.704
Logit PS-LR IPTW	-0.413	0.013	0.083	0.084	11.592	
Logit PS-LR2 IPTW	-0.412	0.012	0.108	0.108	13.734	
SL IPTW	-0.414	0.014	0.091	0.092	12.806	
GBM IPTW	-0.363	0.037	0.084	0.092	18.769	

Table 12. Simulation Results for Scenario S10

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.262	0.138	0.117	0.181	30.981	
Adjusted by Ws	-0.398	0.002	0.073	0.073		
Adjusted by PS-LR	-0.398	0.002	0.073	0.073		
Adjusted by PS-LR2	-0.398	0.002	0.083	0.083		
Adjusted by PS-SL	-0.378	0.022	0.076	0.080		
Adjusted by PS-GBM	-0.565	0.165	0.138	0.215		
Logit PS-LR Matching	-0.403	0.003	0.122	0.122	7.485	6.616
Logit PS-LR2 Matching	-0.380	0.020	0.171	0.172	12.209	10.724
SL Matching	-0.379	0.021	0.116	0.118	10.132	3.690
GBM Matching	-0.541	0.141	0.220	0.261	22.106	62.850
Logit PS-LR IPTW	-0.406	0.006	0.096	0.097	5.200	
Logit PS-LR2 IPTW	-0.394	0.006	0.203	0.203	12.819	
SL IPTW	-0.405	0.005	0.100	0.100	6.312	
GBM IPTW	-0.350	0.050	0.087	0.100	16.990	

Table 13. Simulation Results for Scenario S11

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.296	0.104	0.119	0.159	33.828	
Adjusted by Ws	-0.402	0.002	0.077	0.077		
Adjusted by PS-LR	-0.412	0.012	0.067	0.069		
Adjusted by PS-LR2	-0.414	0.014	0.071	0.073		
Adjusted by PS-SL	-0.382	0.018	0.075	0.077		
Adjusted by PS-GBM	-0.568	0.168	0.124	0.209		
Logit PS-LR Matching	-0.408	0.008	0.105	0.106	20.795	4.874
Logit PS-LR2 Matching	-0.415	0.015	0.110	0.111	21.526	9.002
SL Matching	-0.382	0.018	0.106	0.108	23.829	2.626
GBM Matching	-0.546	0.146	0.175	0.228	32.964	44.528
Logit PS-LR IPTW	-0.410	0.010	0.071	0.071	17.706	
Logit PS-LR2 IPTW	-0.417	0.017	0.100	0.101	19.464	
SL IPTW	-0.415	0.015	0.079	0.081	19.329	
GBM IPTW	-0.382	0.018	0.079	0.081	23.741	

Table 14. Simulation Results for Scenario S12

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.296	0.104	0.119	0.159	33.828	
Adjusted by Ws	-0.402	0.002	0.077	0.077		
Adjusted by PS-LR	-0.402	0.002	0.079	0.079		
Adjusted by PS-LR2	-0.406	0.006	0.093	0.093		
Adjusted by PS-SL	-0.389	0.011	0.084	0.084		
Adjusted by PS-GBM	-0.525	0.125	0.136	0.184		
Logit PS-LR Matching	-0.388	0.012	0.149	0.150	9.541	6.962
Logit PS-LR2 Matching	-0.390	0.010	0.252	0.252	17.336	11.996
SL Matching	-0.375	0.025	0.145	0.147	11.349	3.994
GBM Matching	-0.499	0.099	0.251	0.270	21.839	77.128
Logit PS-LR IPTW	-0.389	0.011	0.133	0.134	8.981	
Logit PS-LR2 IPTW	-0.401	0.001	0.267	0.267	18.799	
SL IPTW	-0.388	0.012	0.128	0.128	9.553	
GBM IPTW	-0.354	0.046	0.092	0.103	20.976	

Table 15. Simulation Results for Scenario S13

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.289	0.111	0.118	0.162	26.744	
Adjusted by Ws	-0.400	0.000	0.069	0.069		
Adjusted by PS-LR	-0.406	0.006	0.068	0.068		
Adjusted by PS-LR2	-0.407	0.007	0.073	0.074		
Adjusted by PS-SL	-0.390	0.010	0.076	0.077		
Adjusted by PS-GBM	-0.546	0.146	0.128	0.194		
Logit PS-LR Matching	-0.411	0.011	0.100	0.100	11.972	4.656
Logit PS-LR2 Matching	-0.416	0.016	0.119	0.120	13.061	10.470
SL Matching	-0.398	0.002	0.112	0.112	14.254	3.908
GBM Matching	-0.513	0.113	0.191	0.221	22.225	46.658
Logit PS-LR IPTW	-0.428	0.028	0.079	0.083	9.281	
Logit PS-LR2 IPTW	-0.410	0.010	0.106	0.106	11.566	
SL IPTW	-0.419	0.019	0.084	0.086	10.665	
GBM IPTW	-0.376	0.024	0.082	0.086	16.181	

Table 16. Simulation Results for Scenario S14

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.289	0.111	0.118	0.162	26.744	
Adjusted by Ws	-0.400	0.000	0.069	0.069		
Adjusted by PS-LR	-0.402	0.002	0.069	0.069		
Adjusted by PS-LR2	-0.402	0.002	0.080	0.080		
Adjusted by PS-SL	-0.404	0.004	0.078	0.078		
Adjusted by PS-GBM	-0.559	0.159	0.146	0.216		
Logit PS-LR Matching	-0.404	0.004	0.110	0.110	6.873	5.438
Logit PS-LR2 Matching	-0.391	0.009	0.155	0.156	10.693	11.578
SL Matching	-0.406	0.006	0.123	0.123	8.576	7.598
GBM Matching	-0.525	0.125	0.227	0.259	21.195	78.858
Logit PS-LR IPTW	-0.425	0.025	0.088	0.091	4.377	
Logit PS-LR2 IPTW	-0.403	0.003	0.174	0.174	11.250	
SL IPTW	-0.408	0.008	0.092	0.093	5.978	
GBM IPTW	-0.362	0.038	0.088	0.096	15.900	

Table 17. Simulation Results for Scenario S15

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.311	0.089	0.118	0.148	27.288	
Adjusted by Ws	-0.398	0.002	0.073	0.073		
Adjusted by PS-LR	-0.408	0.008	0.070	0.071		
Adjusted by PS-LR2	-0.410	0.010	0.072	0.073		
Adjusted by PS-SL	-0.386	0.014	0.077	0.079		
Adjusted by PS-GBM	-0.546	0.146	0.129	0.195		
Logit PS-LR Matching	-0.410	0.010	0.111	0.112	15.593	3.982
Logit PS-LR2 Matching	-0.410	0.010	0.115	0.115	16.300	9.242
SL Matching	-0.389	0.011	0.112	0.113	18.493	2.956
GBM Matching	-0.517	0.117	0.194	0.226	26.381	44.466
Logit PS-LR IPTW	-0.419	0.019	0.077	0.079	12.165	
Logit PS-LR2 IPTW	-0.412	0.012	0.092	0.093	14.240	
SL IPTW	-0.418	0.018	0.082	0.084	14.042	
GBM IPTW	-0.386	0.014	0.082	0.084	18.165	

Table 18. Simulation Results for Scenario S16

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.311	0.089	0.118	0.148	27.288	
Adjusted by Ws	-0.398	0.002	0.073	0.073		
Adjusted by PS-LR	-0.399	0.001	0.073	0.073		
Adjusted by PS-LR2	-0.402	0.002	0.081	0.081		
Adjusted by PS-SL	-0.412	0.012	0.083	0.084		
Adjusted by PS-GBM	-0.523	0.123	0.144	0.189		
Logit PS-LR Matching	-0.409	0.009	0.114	0.114	7.175	4.818
Logit PS-LR2 Matching	-0.388	0.012	0.172	0.172	11.315	9.542
SL Matching	-0.415	0.015	0.141	0.141	10.008	12.694
GBM Matching	-0.506	0.106	0.259	0.280	21.597	100.736
Logit PS-LR IPTW	-0.412	0.012	0.098	0.098	4.786	
Logit PS-LR2 IPTW	-0.396	0.004	0.196	0.196	13.879	
SL IPTW	-0.395	0.005	0.103	0.103	8.127	
GBM IPTW	-0.360	0.040	0.092	0.100	18.286	

Table 19. Simulation Results for Scenario S17

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.521	0.121	0.051	0.131	39.928	
Adjusted by Ws	-0.400	0.000	0.043	0.043		
Adjusted by PS-LR	-0.395	0.005	0.043	0.044		
Adjusted by PS-LR2	-0.400	0.000	0.044	0.044		
Adjusted by PS-SL	-0.409	0.009	0.044	0.045		
Adjusted by PS-GBM	-0.400	0.000	0.044	0.044		
Logit PS-LR Matching	-0.394	0.006	0.058	0.058	12.728	7.578
Logit PS-LR2 Matching	-0.402	0.002	0.053	0.054	11.676	16.344
SL Matching	-0.402	0.002	0.055	0.055	10.659	24.558
GBM Matching	-0.402	0.002	0.053	0.054	11.676	16.344
Logit PS-LR IPTW	-0.367	0.033	0.048	0.058	11.822	
Logit PS-LR2 IPTW	-0.405	0.005	0.053	0.053	8.272	
SL IPTW	-0.401	0.001	0.060	0.060	12.853	
GBM IPTW	-0.408	0.008	0.055	0.055	8.158	

Table 20. Simulation Results for Scenario S18

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.425	0.025	0.088	0.092	22.266	
Adjusted by Ws	-0.400	0.000	0.038	0.038		
Adjusted by PS-LR	-0.400	0.000	0.038	0.038		
Adjusted by PS-LR2	-0.401	0.001	0.042	0.042		
Adjusted by PS-SL	-0.397	0.003	0.047	0.048		
Adjusted by PS-GBM	-0.393	0.007	0.050	0.051		
Logit PS-LR Matching	-0.393	0.007	0.075	0.076	7.165	4.658
Logit PS-LR2 Matching	-0.402	0.002	0.102	0.102	9.646	8.970
SL Matching	-0.392	0.008	0.079	0.079	10.451	2.394
GBM Matching	-0.398	0.002	0.105	0.105	10.557	18.796
Logit PS-LR IPTW	-0.399	0.001	0.044	0.044	2.859	
Logit PS-LR2 IPTW	-0.395	0.005	0.100	0.100	9.330	
SL IPTW	-0.387	0.013	0.053	0.054	4.974	
GBM IPTW	-0.403	0.003	0.048	0.048	5.348	

Table 21. Simulation Results for Scenario S19

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.385	0.015	0.045	0.048	65.771	
Adjusted by Ws	-0.403	0.003	0.039	0.039		
Adjusted by PS-LR	-0.403	0.003	0.040	0.040		
Adjusted by PS-LR2	-0.403	0.003	0.041	0.041		
Adjusted by PS-SL	-0.391	0.009	0.043	0.044		
Adjusted by PS-GBM	-0.403	0.003	0.041	0.041		
Logit PS-LR Matching	-0.404	0.004	0.044	0.044	7.677	0.000
Logit PS-LR2 Matching	-0.404	0.004	0.043	0.044	7.645	0.000
SL Matching	-0.400	0.000	0.054	0.054	12.190	0.150
GBM Matching	-0.404	0.004	0.043	0.044	7.645	0.000
Logit PS-LR IPTW	-0.400	0.000	0.042	0.042	4.402	
Logit PS-LR2 IPTW	-0.404	0.004	0.043	0.044	0.000	
SL IPTW	-0.390	0.010	0.065	0.065	37.962	
GBM IPTW	-0.404	0.004	0.043	0.044	0.000	

Table 22. Simulation Results for Scenario S20

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-0.392	0.008	0.043	0.044	34.746	
Adjusted by Ws	-0.402	0.002	0.036	0.036		
Adjusted by PS-LR	-0.402	0.002	0.036	0.036		
Adjusted by PS-LR2	-0.402	0.002	0.039	0.039		
Adjusted by PS-SL	-0.402	0.002	0.039	0.039		
Adjusted by PS-GBM	-0.402	0.002	0.039	0.039		
Logit PS-LR Matching	-0.403	0.003	0.045	0.045	8.319	0.000
Logit PS-LR2 Matching	-0.403	0.003	0.045	0.045	8.307	0.000
SL Matching	-0.405	0.005	0.057	0.057	11.042	0.466
GBM Matching	-0.403	0.003	0.045	0.045	8.307	0.000
Logit PS-LR IPTW	-0.401	0.001	0.036	0.036	2.498	
Logit PS-LR2 IPTW	-0.403	0.003	0.045	0.045	0.000	
SL IPTW	-0.405	0.005	0.049	0.049	6.974	
GBM IPTW	-0.403	0.003	0.045	0.045	0.000	

Table 23. Simulation Results for Scenario S21

	Estimate	Absolute Bias	Empirical SE	rMSE	ASAM	Discarded
Naive	-19.135	18.735	3.857	19.128	25.964	
Adjusted by Ws	-6.363	5.963	2.303	6.392		
Adjusted by PS-LR	-5.598	5.198	2.135	5.619		
Adjusted by PS-LR2	-3.615	3.215	1.943	3.757		
Adjusted by PS-SL	-7.092	6.692	2.141	7.026		
Adjusted by PS-GBM	7.166	7.566	3.837	8.483		
Logit PS-LR Matching	-6.137	5.737	3.799	6.881	8.141	8.158
Logit PS-LR2 Matching	-5.749	5.349	5.794	7.886	12.495	13.228
SL Matching	-7.539	7.139	3.747	8.062	10.512	3.922
GBM Matching	3.493	3.893	6.172	7.297	21.012	63.006
Logit PS-LR IPTW	-3.405	3.005	8.114	8.653	8.747	
Logit PS-LR2 IPTW	-5.780	5.380	6.555	8.479	12.619	
SL IPTW	-5.940	5.540	3.266	6.431	8.017	
GBM IPTW	-11.215	10.815	2.637	11.132	15.306	

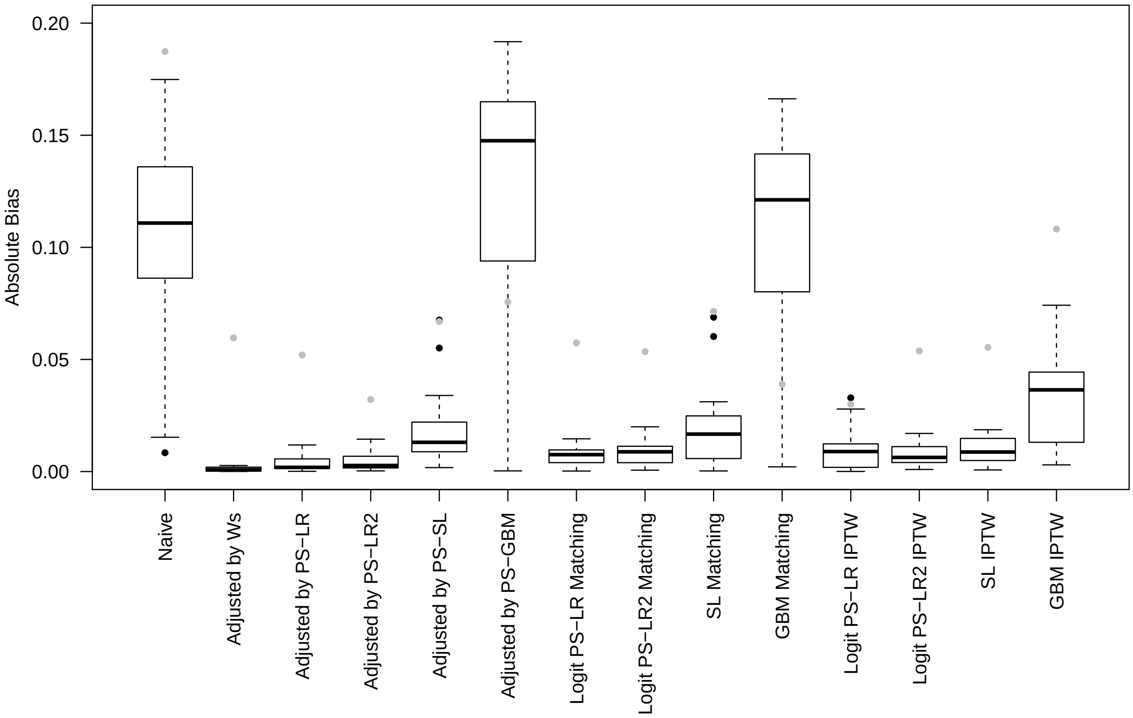


Figure 1. Summary of Simulations Using the Synthetic Data, Aggregated Across 21 Data Generating Scenarios. Absolute Bias by Analysis Method. Grey Points and Number Refer to the Absolute Bias Scaled by 100 for Scenario S21.

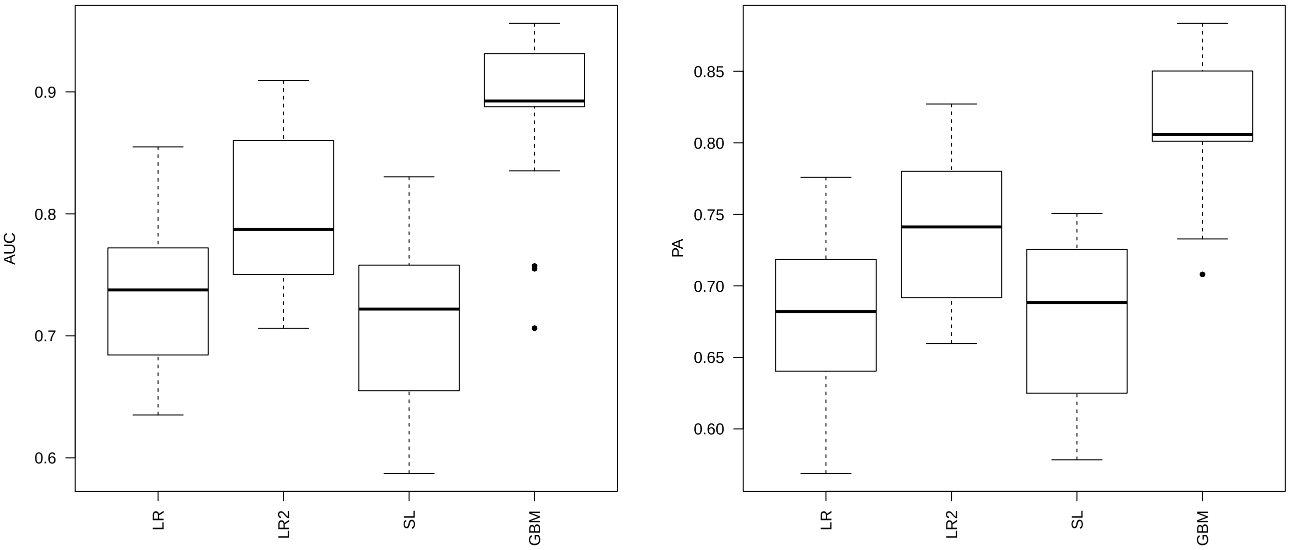


Figure 2. Summary of Simulations Using the Synthetic Data, Aggregated Across 21 Data Generating Scenarios. Plot Showing Area Under the Operating Curve (AUC) Measure and Prediction Accuracy (PA). Left: AUC by PS Models. Right: PA by PS Models

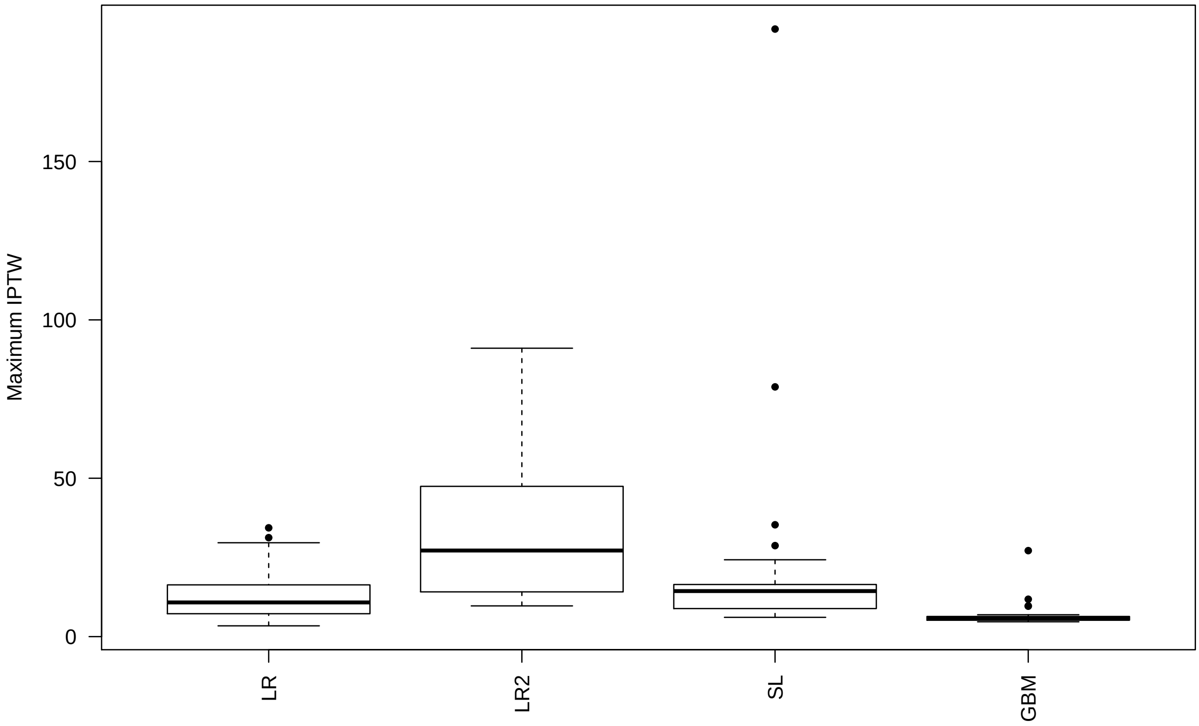


Figure 3. Summary of Simulations Using the Synthetic Data, Aggregated Across 21 Data Generating Scenarios. Maximum of the IPT Weights by Analysis Method.

2. Summary of the Failed Bootstrap Estimates by Method

Table 24. Summary of Bootstraps that did not Converge in the Analysis of First Steps Data (null ATE)

Scenarios	Four Confounders			Eight Confounders		
Sample Size	100	300	500	100	300	500
Naive	0	0	0	0	0	0
Adjusted by Ws	59 (3)	0	0	63 (3)	0	0
Adjusted by PS-LR	59 (3)	0	0	128 (3)	0	0
Adjusted by PS-LR2	59 (3)	0	0	-	81 (3)	0
Adjusted by PS-SL	227 (14)	0	0	226 (16)	2 (1)	0
Adjusted by PS-GBM	0	0	0	85 (3)	0	0
Logit PS-LR Matching	59 (3)	0	0	392 (89)	0	0
Logit PS-LR2 Matching	60 (3)	0	0	-	395 (66)	2 (1)
SL Matching	433 (43)	1 (1)	0	451 (43)	173 (30)	5 (2)
GBM Matching	1 (1)	0	0	473 (38)	32 (2)	0
Logit PS-LR IPTW	60 (3)	0	0	381 (86)	395 (66)	0
Logit PR-LR2 IPTW	69 (3)	0	0	-	400 (66)	0
SL IPTW	392 (42)	15 (1)	0	477 (42)	448 (33)	138 (5)
GBM IPTW	1 (1)	0	0	469 (30)	32 (2)	0

3. Distribution of the Inverse Probability Weights for the Null Treatment Effect from the First Steps Data Analysis

Table 25. Inverse Probability Weights Distribution for the Simulations using First Steps Data

	Min.	Q1	Q2	Mean	Q3	Max.
Scenario A						
Sample Size = 100						
PS-LR	1.0	1.1	1.1	2.0	1.4	21.6
PS-LR2	1.0	1.0	1.1	2.1×10^{12}	1.3	9.0×10^{12}
PS-SL	1.0	1.1	1.1	4.6×10^{6}	1.4	4.6×10^{8}
PS-GBM	1.0	1.1	1.1	1.8	1.4	14.4
Sample Size = 300						
PS-LR	1.1	1.1	1.1	2.0	1.4	17.3
PS-LR2	1.0	1.1	1.1	2.0	1.4	28.4
PS-SL	1.1	1.1	1.1	8.8×10^{2}	1.4	2.6×10^{5}
PS-GBM	1.0	1.1	1.1	1.9	1.4	19.6
Sample Size = 500						
PS-LR	1.1	1.1	1.1	2.0	1.4	15.3
PS-LR2	1.0	1.1	1.1	2.0	1.4	27.0
PS-SL	1.1	1.1	1.1	2.1	1.4	20.0
PS-GBM	1.0	1.1	1.1	2.0	1.4	19.5
Scenario B						
Sample Size = 100						
PS-LR	1.0	1.0	1.1	2.0	1.3	43.0
PS-LR2	1.0	1.0	1.0	6.7×10^{13}	9.0×10^{12}	1.2×10^{15}
PS-SL	1.0	1.0	1.1	2.2×10^{8}	1.5	2.2×10^{10}
PS-GBM	1.0	1.0	1.1	1.4	1.3	7.6
Sample Size = 300						
PS-LR	1.0	1.0	1.1	2.0	1.4	42.5
PS-LR2	1.0	1.0	1.0	1.8×10^{13}	1.2	1.4×10^{14}
PS-SL	1.0	1.0	1.1	2.3×10^{4}	1.4	6.9×10^{6}

PS-GBM	1.0	1.0	1.1	1.5	1.3	20.3
Sample Size = 500						
PS-LR	1.0	1.1	1.1	2.0	1.4	43.1
PS-LR2	1.0	1.0	1.1	4.2×10^{12}	1.3	4.5×10^{13}
PS-SL	1.0	1.0	1.1	3.8×10^{12}	1.4	1.9×10^{14}
PS-GBM	1.0	1.0	1.1	1.6	1.3	26.5
Scenario C						
Sample Size = 100						
PS-LR	1.0	1.1	1.1	2.0	1.4	21.2
PS-LR2	1.0	1.0	1.1	3.3×10^{12}	1.3	2.7×10^{13}
PS-SL	1.0	1.1	1.1	1.9×10^{10}	1.4	1.9×10^{12}
PS-GBM	1.0	1.1	1.2	1.8	1.4	13.8
Sample Size = 300						
PS-LR	1.1	1.1	1.1	2.0	1.4	17.6
PS-LR2	1.0	1.1	1.1	2.0	1.4	26.6
PS-SL	1.1	1.1	1.1	1.8×10^{3}	1.4	5.5×10^{5}
PS-GBM	1.0	1.1	1.1	1.9	1.4	19.5
Sample Size = 500						
PS-LR	1.1	1.1	1.1	2.0	1.4	14.9
PS-LR2	1.0	1.1	1.1	2.0	1.4	24.7
PS-SL	1.1	1.1	1.1	2.1	1.4	19.4
PS-GBM	1.0	1.1	1.1	2.0	1.4	18.5
Scenario D						
Sample Size = 100						
PS-LR	1.0	1.0	1.1	2.0	1.3	43.5
PS-LR2	1.0	1.0	1.8×10^{13}	7.3×10^{13}	2.7×10^{13}	1.1×10^{15}
PS-SL	1.0	1.0	1.1	6.3×10^{11}	1.4	4.0×10^{13}
PS-GBM	1.0	1.0	1.1	1.4	1.3	7.5
Sample Size = 300						
PS-LR	1.0	1.0	1.1	2.0	1.4	43.8
PS-LR2	1.0	1.0	1.0	2.2×10^{13}	1.2	1.7×10^{14}
PS-SL	1.0	1.0	1.1	20.5	1.4	1.8×10^{3}
PS-GBM	1.0	1.0	1.1	1.5	1.3	20.6
Sample Size = 500						
PS-LR	1.0	1.1	1.1	1.9	1.4	42.3
PS-LR2	1.0	1.0	1.1	1.2×10^{12}	1.3	2.7×10^{13}
PS-SL	1.0	1.0	1.1	1.0×10^{3}	1.4	4.9×10^{5}
PS-GBM	1.0	1.0	1.1	1.6	1.3	26.1

4. Results of the Non-Null Treatment Effect from the First Steps Data Analysis

Table 26. Simulation Results for the Simulation using First Steps Data with a Non-null Treatment Effect, with Only Confounders Included in the Estimated Propensity Score and Outcome Models. The ASAM is computed over the confounding variables only.

	Estimate	Absolute Bias	Empirical SE	RMSE	ASAM(%)^1	Discarded
Sample Size = 100						
Naive	110.4	39.6	159.2	164.0	41.1	
Adjusted by Ws	146.3	3.7	151.2	151.2		
Adjusted by PS-LR	146.5	3.5	153.6	153.6		
Adjusted by PS-LR2	148.2	1.8	174.2	174.2		
Adjusted by PS-SL	116.8	33.2	157.4	160.8		
Adjusted by PS-GBM	130.1	19.9	161.6	162.8		
Logit PS-LR Matching	154.4	4.4	205.0	205.1	28.5	11.7
Logit PS-LR2 Matching	155.8	5.8	202.6	202.6	28.0	31.4
SL Matching	119.8	30.2	185.1	187.6	49.6	9.2
GBM Matching	130.4	19.6	203.5	204.5	33.8	16.0
Logit PS-LR IPTW	151.5	1.5	185.7	185.7	18.2	
Logit PS-LR2 IPTW	152.0	2.0	190.5	190.5	24.1	
SL IPTW	141.1	8.9	289.0	289.2	77.5	
GBM IPTW	127.7	22.3	178.2	179.6	22.1	
Sample Size = 300						
Naive	118.4	31.6	95.0	100.1	32.2	
Adjusted by Ws	149.7	0.3	92.7	92.7		
Adjusted by PS-LR	149.2	0.8	93.0	93.0		
Adjusted by PS-LR2	150.5	0.5	95.2	95.2		
Adjusted by PS-SL	133.6	16.4	93.1	94.6		
Adjusted by PS-GBM	149.4	0.6	94.8	94.8		
Logit PS-LR Matching	154.0	4.0	121.1	121.2	18.1	4.4
Logit PS-LR2 Matching	156.7	6.7	126.3	126.4	17.3	16.4
SL Matching	144.4	5.6	122.5	122.6	27.0	7.0
GBM Matching	151.3	1.3	122.0	122.0	18.9	9.1
Logit PS-LR IPTW	151.1	1.1	105.5	105.5	6.5	
Logit PS-LR2 IPTW	158.3	8.3	120.1	120.4	8.6	
SL IPTW	138.1	11.9	110.5	111.2	15.8	
GBM IPTW	148.8	1.2	111.0	111.0	6.1	
Sample Size = 500						
Naive	115.0	35.0	73.9	81.8	30.2	
Adjusted by Ws	149.0	1.0	72.0	72.0		
Adjusted by PS-LR	148.8	1.2	71.9	71.9		
Adjusted by PS-LR2	150.1	0.1	74.2	74.2		
Adjusted by PS-SL	136.1	13.9	71.9	73.3		
Adjusted by PS-GBM	130.7	19.3	73.0	75.5		
Logit PS-LR Matching	154.3	4.3	88.5	88.6	13.8	3.8
Logit PS-LR2 Matching	154.2	4.2	89.5	89.6	14.2	12.7
SL Matching	149.3	0.7	88.2	88.2	21.8	5.1
GBM Matching	127.8	22.2	85.0	87.9	19.1	7.8
Logit PS-LR IPTW	153.2	3.2	78.7	78.7	4.6	
Logit PS-LR2 IPTW	154.1	4.1	85.9	86.0	4.8	
SL IPTW	140.3	9.7	79.1	79.7	10.6	
GBM IPTW	128.7	21.3	81.7	84.4	8.8	

[1] Balance is checked for covariates child's sex, mother's race, parity and smoking status.

Table 27. Simulation Results for the Simulation using First Steps Data with a Non-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.

	Estimate	Absolute Bias	Empirical SE	RMSE	ASAM(%)[1]	Discarded
Sample Size = 100[2]						
Naive	86.1	63.9	179.2	190.3	56.4	
Adjusted by Ws	150.2	0.2	167.6	167.6		
Adjusted by PS-LR	151.0	1.0	176.7	176.7		
Adjusted by PS-SL	114.4	35.6	172.1	175.7		
Adjusted by PS-GBM	197.0	47.0	242.2	246.7		
Logit PS-LR Matching	157.8	7.8	300.3	300.4	36.1	21.8
SL Matching	109.3	40.7	225.4	229.0	55.7	11.8
GBM Matching	178.2	28.2	338.5	339.7	54.0	65.8
Logit PS-LR IPTW	147.9	2.1	259.0	259.0	32.3	
SL IPTW	121.7	28.3	321.6	322.8	137.0	
GBM IPTW	118.2	31.8	179.4	182.2	37.6	
Sample Size = 300						
Naive	96.5	53.5	101.9	115.1	49.7	
Adjusted by Ws	148.8	1.2	97.6	97.6		
Adjusted by PS-LR	146.5	3.5	97.5	97.6		
Adjusted by PS-LR2	146.9	3.1	122.8	122.8		
Adjusted by PS-SL	134.9	15.1	98.0	99.2		
Adjusted by PS-GBM	176.2	26.2	125.2	127.9		
Logit PS-LR Matching	147.7	2.3	177.9	177.9	19.1	5.1
Logit PS-LR2 Matching	146.3	3.7	297.8	297.8	36.5	17.2
SL Matching	138.1	11.9	154.3	154.8	25.1	4.9
GBM Matching	161.6	11.6	250.2	250.4	32.3	33.2
Logit PS-LR IPTW	145.3	4.7	150.5	150.6	14.1	
Logit PS-LR2 IPTW	141.7	8.3	253.4	253.5	33.3	
SL IPTW	148.4	1.6	181.2	181.2	22.2	
GBM IPTW	134.8	15.2	125.8	126.7	20.4	
Sample Size = 500						
Naive	91.9	58.1	74.0	94.1	48.6	
Adjusted by Ws	145.6	4.4	71.1	71.3		
Adjusted by PS-LR	142.3	7.7	71.5	72.0		
Adjusted by PS-LR2	145.8	4.2	80.3	80.5		
Adjusted by PS-SL	138.9	11.1	72.7	73.5		
Adjusted by PS-GBM	170.6	20.6	85.9	88.3		
Logit PS-LR Matching	140.2	9.8	126.0	126.4	15.4	2.6
Logit PS-LR2 Matching	148.4	1.6	184.6	184.6	21.3	9.1
SL Matching	144.8	5.2	124.1	124.2	19.4	2.7
GBM Matching	152.9	2.9	205.3	205.3	27.0	17.5
Logit PS-LR IPTW	142.1	7.9	97.0	97.3	10.3	
Logit PS-LR2 IPTW	153.5	3.5	196.5	196.5	25.7	
SL IPTW	153.9	3.9	185.1	185.2	23.3	
GBM IPTW	131.8	18.2	95.8	97.5	16.7	

[1] Balance is checked for covariates child's sex, mother's age, race, parity, marital status, smoking status, weight prior to pregnancy and education level

[2] Analysis for propensity score model with main effects plus all two way interaction terms (propensity score-LR2) was very unstable with sample size 100. Therefore, PS-LR2 estimates were not calculated.

5. Code

In an additional supplementary file, we provide sample code for the plasmode simulation using the First Steps database. The code provided are for the scenario with a null treatment effect and four confounders, with a sample size of 500.

References

Diaz, I. and Kelly, J. (2016). To balance or not to balance? http://www.unofficialgoogledatascience.com/2016/06/to-balance-or-not-to-balance.html.

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

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

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, 17(6):546–555.