Stacked Ensemble Learning for Propensity Score Methods in Observational Studies

Maximilian Autenrieth
Imperial College London
m.autenrieth@web.de

Richard A. Levine
San Diego State University
rlevine@sdsu.edu

Juanjuan Fan
San Diego State University
jjfan@sdsu.edu

Maureen A. Guarcello
San Diego State University
mguarcello@sdsu.edu

Propensity score methods account for selection bias in observational studies. However, the consistency of the propensity score estimators strongly depends on a correct specification of the propensity score model. Logistic regression and, with increasing popularity, machine learning tools are used to estimate propensity scores. We introduce a stacked generalization ensemble learning approach to improve propensity score estimation by fitting a meta learner on the predictions of a suitable set of diverse base learners. We perform a comprehensive Monte Carlo simulation study, implementing a broad range of scenarios that mimic characteristics of typical data sets in educational studies. The population average treatment effect is estimated using the propensity score in Inverse Probability of Treatment Weighting. Our proposed stacked ensembles, especially using gradient boosting machines as a meta learner trained on a set of 12 base learner predictions, led to superior reduction of bias compared to the current state-of-the-art in propensity score estimation. Further, our simulations imply that commonly used balance measures (averaged standardized absolute mean differences) might be misleading as propensity score model selection criteria. We apply our proposed model – which we call GBM-Stack – to assess the population average treatment effect of a Supplemental Instruction (SI) program in an introductory psychology (PSY 101) course at San Diego State University. Our analysis provides evidence that moving the whole population to SI attendance would on average lead to 1.69 times higher odds to pass the PSY 101 class compared to not offering SI, with a 95% bootstrap confidence interval of (1.31, 2.20).

Keywords: educational data mining, machine learning, ensemble learning, stacked generalization, propensity score estimation, causal inference

1. Introduction and Related Literature

In educational studies, randomized controlled trials are often not feasible since treatments cannot be randomly assigned. For instance, most colleges provide student help centers, subject-specific tutoring centers, stretch courses, or supplemental online material to support students to succeed in their programs. Usually, it is the students' free choice whether they make use of such offerings (treatment) or not. Hence, attendance/participation can be observed but not randomly assigned. In this case, it is not straightforward to draw a conclusion about the effect of a treatment if systematic differences between treated and control observations are present. For instance, if only well-performing students attend a voluntary add-on course, one would highly overestimate the effect of the course by just taking the differences of the outcomes in the treated and control groups. Generally speaking, confounding covariates, associated with both treatment and outcome, could lead to severe bias in average treatment effect estimates.

The introduction of the propensity score methodology by Rosenbaum and Rubin (1983) has been groundbreaking in the causal inference literature to address this issue, with rapidly growing attention in recent years. Rosenbaum and Rubin (1983) introduced the propensity score as the probability of treatment assignment given the observed covariates. They showed that conditional on the propensity scores, treated and untreated observations have the same covariate distributions, under certain assumptions. Therefore, a correctly specified propensity score model in an observational study allows one to replicate characteristics of randomized trials, in particular, to obtain unbiased treatment effect estimates. One widely applied propensity score method for estimation of unbiased population average treatment effects (ATE) is inverse probability of treatment weighting (IPTW). Weights are assigned to the observations in the original sample, such that covariates of treated and untreated observations are balanced in the synthetic weighted sample. In IPTW, the estimated propensity scores are directly incorporated in the sample weights, which are given by the inverse probability of the subjects' treatment status. Thus, to ensure consistency of the average treatment effect estimation, correct specification of the propensity scores is essential. Logistic regression is mainly used as a method to estimate the propensity scores. Recently the use of machine learning algorithms, especially generalized boosted models (GBM), neural networks (NNET), and the Superlearner (SL), showed improvement over logistic regression models (Setoguchi et al., 2008; Lee et al., 2010; Westreich et al., 2010; Pirracchio et al., 2014).

We propose the use of a stacked generalization ensemble learning approach, tailored for the estimation of propensity scores used in IPTW, to reduce bias in population average treatment effect estimation. We implement a comprehensive Monte Carlo simulation study to assess the performance of our proposed models. Our simulation study aims to cover a broad range of realistic data sets in educational studies, being composed of three different data sizes, eight scenarios with different degrees of confounding, and a range of six different treatment effect magnitudes. Finally, we apply our best model, which we call GBM-Stack, on a real-life educational data set.

Stacked generalization is an ensemble learning approach introduced by Wolpert (1992). Roughly summarized, the approach incorporates a set of several diverse base learners (e.g., stable and unstable, parametric and non-parametric models) which are trained on the data on a first level. On a second level, the base learner outcome predictions are put together to build a new data set. This data set is then used as training input for a meta learner which generalizes the base learner predictions to provide the final model predictions. Thereby, the meta learner might learn if the data has been learned properly by the base learners, and might be able to identify if any particular base learners classify certain regions of the features well or not (Polikar, 2007). Diversity of the base learners leads to a broader "coverage" of regions where features are well classified by at least one base learner. Intuitively, the stacked ensemble could be considered as an imitation of human decision-making (Polikar, 2006). For important decisions, various different "experts" (base learners) are asked about their opinion (predictions) of a subject (data). These opinions (predictions) are then considered to make the final decision (meta learner prediction). It has been shown that stacked generalization ensembles out-perform their underlying base learners in regression and classification tasks (Breiman, 1996; Leblanc and Tibshirani, 1996; Ting and Witten, 1999).

We suppose that, especially in propensity score estimation, the use of stacked ensemble learners is of great value. In supervised machine learning tasks, one can easily assess model performance by evaluating suitable performance metrics on a validation set, e.g., RMSE for regression tasks, and ROC or accuracy for classification tasks. In propensity score estimation tasks, performance evaluation is not as clear, since the target error, in this case, the error of the population average treatment effect estimation, is not directly accessible. It has been indicated that a good model performance, regarding the prediction accuracy of treatment assignment, does not necessarily lead to a good performing propensity score model (Westreich et al., 2011; Griffin et al., 2017). In the propensity score literature, balance measures are introduced to assess model performance (e.g., Austin and Stuart 2015), since balanced covariates indicate that there are no systematic differences between treatment and control groups. However, in the literature there is no overall agreement or criterion for a well-balanced data set. Our simulations suggest that averaged standardized absolute mean differences (ASAM), the currently most popular balance measure, might not be a suitable propensity score model selection criterion. One can thus not directly assess in which situation a specific "expert" performs best. This strengthens our motivation for the stacked generalization approach, which performs well on a broad range of scenarios in our simulation study by incorporating a variety of "expert" models in the base learner set. Precisely, our simulations suggest that using logistic regression or GBM as meta learner trained on the predictions of twelve diverse base learners leads to superior bias reduction in the ATE estimation compared to current state-of-the-art propensity score estimation methods, described briefly in the following.

Past work has put forth vast effort to obtain consistent propensity score estimators (Setoguchi et al., 2008; Lee et al., 2010; Westreich et al., 2010). Lately, generalized boosted models (GBM) implemented in the `twang-package` (Ridgeway et al., 2014), henceforth denoted as Twang-GBM, which aims to maximize covariate balance, and the Superlearner (SL; Pirracchio et al. 2014; Pirracchio and Carone 2018) showed superior performance in propensity score estimation. The Twang-GBM model is a boosted ensemble (we refer to Ridgeway et al. 2014 and Friedman 2001 for more details) and not considered as a stacked ensemble in the sense of Wolpert (1992). Technically, Superlearner can be considered as a stacked ensemble learner, since it employs a constrained linear combination of different base learner predictions, minimizing a given loss function. However, no meta learners, such as GBM or logistic regression, are trained on the base learner set. To our knowledge, there are no other stacked generalization approaches using different meta learners trained on a suite of base learners in the statistical literature applied to propensity score estimation.

Observational studies are prolific in a broad range of educational studies, and propensity score methods are common approaches for analyzing such data. To provide a method that can be widely applied, in our Monte Carlo simulation study we aim to cover a broad range of realistic observational, educational data. To give a brief overview, there have been several student success studies assessing the effect of supplemental course, tool, or seminar offerings with voluntary student attendance (Clark and Cundiff, 2011; Jiang and McComas, 2015; Feild et al., 2016; Guarcello et al., 2017). Propensity scores have been applied to account for the confounding introduced through the voluntary attendance in the programs. A path of research with high attention in this area is for instance the assessment of special education and inclusive programs for disadvantaged students or students with disabilities (Shapiro and Trevino, 2004; Morgan et al., 2010; Sullivan and Field, 2013; Rojewski et al., 2015; Bakker et al., 2020). Others employ propensity scores to assess the influence of education types on social, civic, and political participation (Kam and Palmer, 2008; Brand and Xie, 2010; Kim and Clark, 2013). Propensity scores have further been used to evaluate the impact of different education programs on later career success (Titus, 2007), to assess the impact of teacher encouragement on enrollment in advanced high schools or college (Alcott, 2017), or to evaluate a possible positive effect on recidivism through college education for prisoners (Kim and Clark, 2013). We emphasize that the summary of examples provided is by no means exhaustive, but it serves as an illustration of the potential widespread applicability of our proposed stacked generalization propensity score estimators, by providing superior bias reduction in average treatment effect estimation.

In Section 5, we add to the above described educational propensity score literature by applying our best propensity score estimator, namely GBM-Stack, on a real-world educational data set. Specifically, we investigate the effect of a Supplemental Instruction program (SI) on student success for psychology majors' compulsory introduction course (PSY 101) at San Diego State University. PSY 101 has been identified as a bottleneck course at SDSU with a high failure rate of 15%. A high course fail rate jeopardizes a student's four-year graduation plan. SI is a voluntary, weekly add-on program, originally created at the University of Missouri Kansas City (UMKC, 2018), and installed to improve student success. In our context, we assess student success by the performance (pass/fail) of students in the PSY 101 course, given data of 2,173 students collected in 2015, 2016, and 2017. The estimated propensity scores are used in IPTW to estimate the average treatment effect of SI in the whole PSY 101 student population. Our analysis suggests that moving the whole PSY 101 student population to SI attendance would on average lead to 1.69 times higher odds to pass the PSY 101 class compared to not offering SI. We obtain a 95% bootstrap confidence interval of (1.31, 2.20). The analysis of this real-world data serves as both, an illustration of our proposed stacked generalization propensity score estimation approach, and also as evidence for the positive effect of SI on the PSY 101 student population, which could have important implications on the funding and extensions of the SI program.

The structure of our paper is as follows. In Section 2, we formally introduce the propensity score methodology and substantiate the construction of our stacking approach for propensity score estimation. In Section 3, we introduce our simulation study. An extensive Monte Carlo simulation study, mimicking realistic scenarios in educational studies, is conducted to investigate the performance of our proposed stacked ensembles. In Section 4, we present the results of our simulation study. We use the estimated propensity scores in IPTW to estimate population average treatment effects (ATE), and compare the performance of our proposed ensemble learners to the current state-of-the-art models according to bias, mean squared error, and variance of the point estimate. We assess the weights used in IPTW and the ASAM, the most commonly applied measure of balance, to check the balance of our models. We critically discuss the relation of the ASAM with ATE estimation errors. In Section 5, we apply our proposed model, GBM-Stack, on recent student data to estimate the population average treatment effect of the SI program on student success in the PSY 101 course at San Diego State University. A final discussion of our findings appears in Section 6, with limitations and possible future work in Section 7.

2. Analytical methods

2.1. Propensity score

The aim of the propensity score method is to generate balanced covariates, between treated and untreated observations, to control for confounding caused by selection bias in observational studies. Rosenbaum and Rubin (1983) introduce the propensity score as the probability of treatment assignment given a vector of observed covariates, denoted by

$$e(X) = P(Z = 1|X), \tag{1}$$

where Z is a binary variable indicating that an observation is exposed to the treatment, i.e. $(Z = 1)$, or not exposed to the treatment, i.e. $(Z = 0)$. X is the vector of observed covariates. As a fundamental concept, to formulate conditions for non-confounding, Rosenbaum and Rubin (1983) define treatment allocation as strongly ignorable if

$$\text{(i)}\;\; (Y^{(1)}, Y^{(0)}) \perp\!\!\!\perp Z|X \qquad \text{and} \qquad \text{(ii)}\;\; 0 < e(X) < 1. \tag{2}$$

Here, (i) means that treatment assignment Z and the potential outcome $(Y^{(1)},\ Y^{(0)})$ are conditionally independent given the covariates X. Note that the potential outcomes are the outcome values that are possible for a certain object. The potential outcomes are never observed at once, since the actual observed outcome Y takes on one of the values, depending on the treatment status of the object. For instance, if an object is treated, then $Y = Y^{(1)}$ is observed. (ii) means that the probability for either treatment or control group allocation has to be non-zero. Rosenbaum and Rubin (1983) demonstrate that the propensity score is a balancing score. If treatment allocation is strongly ignorable, conditional on the propensity score an unbiased average treatment effect (ATE) may be obtained. This implies that a correct specification of the propensity scores allows us to imitate characteristics of randomized control trials in situations when random treatment allocation is unfeasible (Austin, 2011; Austin and Stuart, 2015). Condition (i) in (2) can be violated if confounders connected to both treatment and outcome remain unmeasured (Austin, 2011).

Mainly four propensity score methods have been proposed in the statistics literature: stratification or subclassification on the propensity score, matching on the propensity score, covariate adjustment using the propensity score, and inverse probability of treatment weighting (IPTW) (Rosenbaum and Rubin, 1983; Rosenbaum and Rubin, 1984; Rosenbaum, 1987). Matching on the propensity score and IPTW have seen increasing popularity in the statistics literature in recent years due to their superior ability to reduce bias. In stratification, or matching on propensity scores, the estimated propensity scores are used to group samples in strata and pairs with desirably similar covariate distributions. In IPTW and covariate adjustment approaches, the actual propensity score estimates are incorporated in the treatment effect estimations. We refer to Austin (2011), Austin and Stuart (2015) and Williamson et al. (2012) for a detailed description and comparison. In this study, we focus on the performance of our proposed ensemble learner to estimate propensity scores used for IPTW.

By utilizing IPTW to estimate the population average treatment effect (ATE), a synthetic sample is created by weighting each observation in the sample, using the inverse probability of the subject's treatment status, with the aim to balance covariates between the treated and untreated group in the synthetic weighted sample. The weights to estimate the ATE are defined by

$$w_{ATE} = \frac{Z}{e(X)} + \frac{1 - Z}{1 - e(X)}. \tag{3}$$

Thus, the weight for each subject directly incorporates the estimated propensity scores and can be described as the inverse probability of the subject's treatment status. It has been shown that

IPTW is advantageous to estimate the ATE for the whole population, whereas matching on the propensity score estimates the average treatment effect in the actual treated group (ATT; Imbens 2004; Austin 2010). The ATT can also be obtained through IPTW by multiplying the ATE weights with the propensity score $e(X)$ (Appendix, Formula 12). In our study, we focus on the estimation of the average treatment effect in the population (ATE). Lunceford and Davidian (2004) show that under the assumption of a correctly specified propensity score model, a consistent estimator of the average treatment effect for a sample of n subjects is given by

$$ATE = \left(\sum_{i=1}^{n} \frac{Z_i}{e_i(X)}\right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{e_i(X)} - \left(\sum_{i=1}^{n} \frac{1 - Z_i}{1 - e_i(X)}\right)^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i)Y_i}{1 - e_i(X)} =: \hat{\mu}_1 - \hat{\mu}_2 =: \hat{\gamma}_{RD}, \qquad (4)$$

where Y_i denotes the outcome variable, Z_i the treatment status, and $e_i(X)$ the propensity score with observed covariates X for subject $i \in \{1, \ldots, n\}$, respectively. Obtaining an unbiased ATE estimate is the main objective of the IPTW approach in our study.

A crucial step in the IPTW procedure is to assess the balance of the covariates between treated and untreated in the synthetically weighted sample (Austin and Stuart, 2015). Imbalance in the weighted sample can for instance indicate a misspecified propensity score model or a violation of the strong ignorability assumption, potentially leading to severe bias in the treatment effect estimate (Formula 4). A predominantly used measure to assess balance in the propensity score literature is the standardized mean difference (smd) of the covariates (Austin and Stuart, 2015). The standardized mean difference measures the difference of the average value of a covariate in the treatment and control groups, scaled by the pooled standard deviation of the measured covariate. Precisely, let $\bar{x}_{treatment}$ and $\bar{x}_{control}$ be the sample mean and $s^2_{treatment}$, $s^2_{control}$ be the sample variance of a covariate in the treatment and control group, respectively. Then,

$$\text{smd} = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}. \qquad (5)$$

To get a summary measure across all covariates, the average of the standardized absolute mean differences (ASAM) over all measured covariates is frequently considered (Lee et al., 2010; Pirracchio et al., 2014; Pirracchio et al., 2016). There are varying recommendations regarding the exact use of standardized mean differences, particularly the ASAM, as a balance measure in propensity score studies. We provide a short summary of the literature in the Appendix (Section 10). In our simulation study, we assess if the ASAM is a suitable model selection criterion given a set of different propensity score estimation models on a particular data set.

2.2. Ensemble Learning - Stacked Generalization for Propensity Score Estimation

We construct our models as a variation of the stacked generalization approach introduced by Wolpert (1992). In brief, a set of L base learners are trained on the learning data set, called level-0-data, and predictions for each observation are obtained through cross-validation. Instead of selecting the base learners with the best performance, the cross-validation predictions of all base learners are merged to a second data set, the level-1-data. Thus, for each observation in the actual data set, the level-1-data contains the L base learner predictions of its class label (treatment status). This L-dimensional set of predictions, level-1-data, serves as an input for a meta learner. Eventually, the output of the meta learner, trained on the level-1-data, provides the

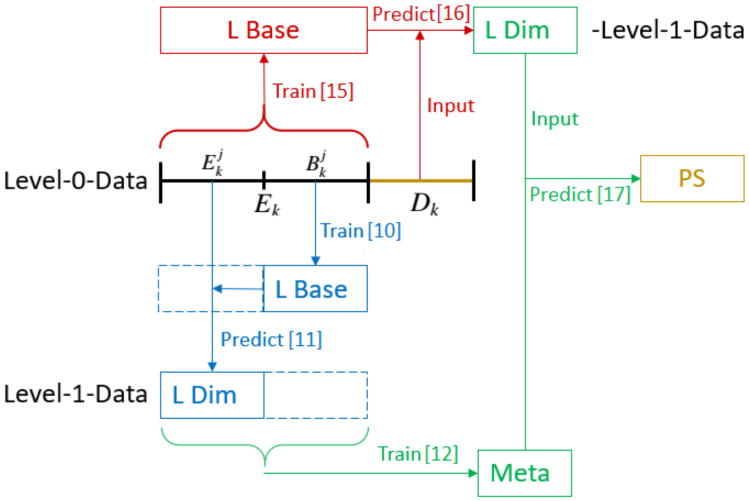

Figure 1: Simplified stacked generalization scheme for propensity score (PS) estimation. The scheme sketches one outer loop k from Algorithm 1 with one inner loop j, to predict the propensity scores for observations in D_k. Some of the steps in Algorithm 1 are marked with square brackets.

ultimate classification or regression predictions. In our case, the class (treatment) probability output of the meta learner provides an estimate for the propensity score of each observation.

In Algorithm 1, we present pseudo-code for a precise description of our proposed stacked generalization approach. Figure 1 serves as a simplified illustration of the approach, representing one outer cross-validation loop $k \in \{1, \dots, K\}$ and one inner loop $j \in \{1, \dots, J\}$, to get propensity score predictions for observations in D_k. In Figure 1, the case $K = 3$ and $J = 2$ of Algorithm 1 is considered for simplicity.

Some technical details have to be considered for a successful design of a stacked ensemble learner in our context. Nested, stratified cross-validation is applied to construct the training and validation sets for the level-0-data and level-1-data. The outer K-fold cross-validation loop (steps 3, 4, and 12 to 17) in Algorithm 1 splits the data D into K disjoint and equal-sized subsets, stratified by the binary response variable. In each loop k, we take one fold D_k out as the ensemble validation set (step 4). The remaining $(k - 1)$ folds build the ensemble training set E_k.

In the second loop (steps 6 to 11) each ensemble training set E_k is further stratified divided into J disjoint and equal-sized folds. Again, in each inner loop, one fold j builds the base learner validation set E_k^j, and the remaining $(j - 1)$ folds build the base learner training set B_k^j (step 7). The preprocessing (steps 8 and 9) ensures that no information from the validation set is used in the training set. This is essential, since leaking information from the validation set to preprocess the training set (e.g., imputing, scaling) could lead to over-optimistic estimation errors on the validation set and therefore to biased level-1-data. Step 9 ensures that the validation set receives the same preprocessing method as the training set. Subsequently, in each inner loop j, each of the L base learners are fit on the preprocessed base training set B_k^j (step 10). The L fitted base learners are then used to predict the class probabilities for each observation on the base validation set E_k^j (step 11). After J inner loops (step 6 - 11), we obtain the L-dimensional

Algorithm 1 Stacked Generalization for Propensity Score Estimation

```
 1: Choose a suitable set of $L$ base learners and one meta learner.
 2: Randomly partition the data $D$ into $K$ disjoint and equal sized subsets $D_1, \dots, D_K$.
 3: Fix the cross-validation counter $k = 1$. (Note that $k \in \{1, \dots, K\}$).
 4: Label the subset $D_k$ as validation set, and $E_k := D - D_k$ as the ensemble training set.
 5: Randomly partition the ensemble training set $E_k$ into $J$ disjoint and equal sized subsets
    $E_k^1, \dots, E_k^J$.
 6: for $j \in \{1, \dots, J\}$ do
 7:     Label $E_k^j$ as validation set and $B_k^j := E_k - E_k^j$ as base learner training set;
 8:     Impute missing values and Scale numerical variables on the training set $B_k^j$;
 9:     Apply the imputing and scaling method from 8 on the validation set $E_k^j$;
10:     Fit each of the L base learners on the training set $B_k^j$;
11:     Predict each observation in the validation set $E_k^j$ with each of the $L$ base learners.
12: Fit the meta learner on the level-1-data of $E_k$, obtained through $J$ inner loops (steps 6 - 11).
13: Impute missing values and Scale numerical variables on the ensemble training set $E_k$.
14: Apply the imputing and scaling method from 13 on the validation set $D_k$.
15: Fit each of the L base learners on the ensemble training set $E_k$.
16: Predict each observation in $D_k$ with each of the $L$ base learners from 15.
    $\Rightarrow$ Obtain L-dimensional level-1-data on the validation set $D_k$.
17: Predict fitted meta learner from 12 on level-1-data from 16 to obtain estimated propensity
    scores for observations in $D_k$.
18: Increment $k$ by one.
19: Repeat steps 4 to 18 until $k > K$.
```

level-1-data for each observation in the ensemble training set E_k as input for our meta learner (step 12). Thus, instead of using the training set predictions of the base learners, the inner loop ensures that just validation set predictions are used as level-1-data and meta learner input (Ting and Witten, 1999; Alpaydin, 2014). No scaling is needed for the level-1-data (Alpaydin, 2014, chap. 17).

In each outer loop k, the meta learner is fitted on the obtained L-dimensional level-1-data of the ensemble training set E_k (step 12). Recall that the level-1-data is obtained for each data point in E_k through steps 6 - 11, but not for D_k. To get level-1-data for the observations in the ensemble validation set D_k, in step 15, the L base learners are fit to the level-0-data of the ensemble training set E_k and predict on each observation in the ensemble validation set D_k (step 16). To perform steps 15 and 16, the level-0-data of E_k and D_k has to be preprocessed (steps 13 - 14), again without leakage of information from the validation set D_k. Note that E_k was already stepwise preprocessed through the inner loop (steps 8 - 9). Generally, these preprocessing steps could also be applied to D_k which would make step 13 redundant. However, in step 8 the preprocessing is done on subsets of E_k. To include all training set information in the preprocessing, we again preprocess the complete E_k (step 13) and apply it on D_k in step 14.

In step 17, the fitted meta learner from step 12 is used to obtain predictions (propensity scores) for the observations in the validation set D_k. Therefore, after K loops we get the stacked ensemble predicted class probabilities, more precisely the propensity scores, for each observation in the data D.

Note that in steps 10, 12, and 15 of Algorithm 1, additional cross-validation loops or bootstrapping can be executed in each model fitting process to select optimal tuning parameter configurations for each base learner and the meta learner. In steps 8-9 and 13-14 further preprocessing functions (e.g., principal component analysis or Box-Cox transformations) can be incorporated if necessary for the underlying data.

2.3. Learner selection and computation

The choice of base and meta learners is modular: generally, any model can be used. Nevertheless, to improve performance, the learners should fulfill some desired properties. The base learners have to be both accurate and diverse (Alpaydin, 2014; Sesmero et al., 2015). A higher diversity in the base learners' predictions leads to more information in the level-1-data set. Loosely speaking, the meta learner needs various opinions that it can rely on. We found it advantageous to include a mixture of both stable and unstable, as well as parametric and non-parametric base learners. Without a certain performance level regarding a chosen evaluation measure, e.g., classification accuracy, ROC curve, kappa, or log loss (Kuhn, 2008), the base learners do not contribute information or may even contribute misleading information to the ensemble. However, it has been shown that the predictive accuracy of stacked ensembles can be increased in a standard classification task by even removing the most accurate base learners from the level-1-data in favor of less correlation, e.g. (Doumpos and Zopounidis, 2007). This trade-off has to be considered when building a suitable set of base learners. In our simulations, we chose a wide-ranging set of $L = 12$ base learners. To maximize diversity, we included both parametric models (e.g., logistic regression and naive Bayes) and non-parametric models (e.g. K-nearest neighbors classification, decision trees, and support-vector-machines). Table A.4 in the Appendix includes short descriptions of the included models.

The aim of the meta learner is to balance the base learners' biases by learning from the base learners' errors (Alpaydin, 2014; Sesmero et al., 2015). A desired property of the meta learner is the ability to generalize and capture the structure of the level-1-data to minimize a selected error metric. To balance the errors of the base learners it is crucial that the meta learner is trained on base learner predictions outside of the base learner training set. Otherwise, the meta learner just strengthens the predictions of overfitted base learners on the training set (Ting and Witten, 1999; Alpaydin, 2014). Ting and Witten (1999) suggest to build the level-1-data with the predicted class probabilities instead of using the binary classifications. In the choice of the meta learner, the model's ability to predict class probabilities with high accuracy was a decisive factor. It is further desired that probabilities are not overestimated towards the edges (zero and one), since the estimated class (treatment) probabilities of the meta learner, specifically the propensity scores, are directly used to construct sample weights (Section 2.1, Formula 3). Wrongly classifying an object with high confidence (estimated probabilities towards zero or one) leads to extreme weights, which strongly influence the estimation of the average treatment effect (Formula 4). Erroneously high weights can lead to severe errors in the ATE estimation and are thus not desirable. We implemented two different meta learners: logistic regression (LR) and gradient boosting machine (GBM). Due to its probabilistic classification approach and its widespread popularity, logistic regression is a commonly used method in propensity score estimation (Austin and Stuart, 2015; Rubin, 2004). The main issue in the application of logistic regression models is misspecification of interaction terms or nonlinear relationships in the data. Thus, without deep knowledge of the data, logistic regression might result in poor model fit and bias in the treatment effect estimate (Drake, 1993). We presume that stacking might boost the performance of logistic regression by handling the structure of the data at the base learner level. Gradient Boosting Machines have gained more attention in the propensity score literature in the last few years. (Parast et al., 2017; Griffin et al., 2017; Stone and Tang, 2013; McCaffrey et al., 2004) suggest GBM in IPTW due to low treatment effect estimation errors in scenarios with unknown, non-linear functional forms in the association of the covariates with the treatment selection.

To our knowledge, there is no overall agreement in favor of a suitable loss function for propensity score model fitting and tuning parameter selection. Classification performance metrics, e.g. L2 squared error or AUC, are applied (Lee et al., 2011; Pirracchio et al., 2014) as well as covariate balance measures, such as the average or maximum of absolute standardized mean differences or the Kolmogorov-Smirnov statistic (Pirracchio and Carone, 2018; Ridgeway et al., 2014). The optimization regarding covariate balance is computationally more expensive than the optimization with respect to the more conventional classification performance metrics. Mainly due to computational reasons, we found the default "off-the-shelf" model training and tuning parameter selection in `caret` as adequate in this paper. Thus, in steps 10, 12, and 15 of Algorithm 1, each base and meta learner was evaluated using 10-fold cross-validation with three different values for each tuning parameter, maximizing the area under the ROC curve. Especially at the meta learner level, balancing loss functions might be incorporated in future investigations.

The nested cross-validation loops in Algorithm 1 can be computationally expensive. However, the modularity of the stacked ensemble building leads to several opportunities to reduce computational cost by parallel processing (Microsoft Corporation and Weston, 2017), since base learners can be trained independently. As a benchmark, using parallel processing on a MacBook Pro with a 2.3 GHz 8-Core i9 processor, in our data analysis in Section 5 with a data set of 2,173 observations and 33 covariates, we could reduce the complete base and meta learner training of our GBM-Stack model to approximately 6 minutes. All base and meta learner models were implemented with the `caret` package (Kuhn, 2008) in R version 3.6.2 (R Core Team, 2018).

3. Monte Carlo simulation study

3.1. Data generation

An important factor in creating our simulation study was to generate data sets consistent with a broad range of applications in educational studies. For this reason, we incorporated characteristics of the student data set that we analyze in Section 5, as well as features and findings of previous student success studies (Guarcello et al., 2017; Beemer et al., 2018; Pelaez et al., 2019). Generally, we suppose that a relatively high proportion of categorical, particularly binary variables and some continuous variables occur in data sets with student demographics and performance covariates. Further, three types of covariates are present, differing in their associations with the treatment and the outcome – confounders, associated with both treatment and outcome; instrumental variables, just associated with the treatment; and covariates just associated with the outcome. Besides, we assume different complex associations of the covariates with the treatment variable (Section 3.2). In the analysis of real-world educational data, expert knowledge or previous studies could help to partly identify the type of the covariates. However, we find it most realistic that the associations of the covariates with the treatment and outcome variable are unknown. We found the simulation study introduced by Setoguchi et al. (2008) as an appropriate basis for our purpose. We adjusted the scenarios (A-G) in Setoguchi et al. (2008). These seven scenarios incorporate various non-linear and non-additive associations between the covariates and the probability to receive the treatment (propensity score). We added one model (Scenario H), with severe non-additivity, that we found realistic in our application.

For each scenario three different sample sizes containing $n_{total} \in \{1000, 2000, \text{ and } 3000\}$ observations were implemented, and $m = 1000$ data sets were generated for each data size. In each data set, a binary treatment variable Z, with $P(Z) \in [0.27, 0.45]$, and binary outcome Y, with $P(Y) \in [0.10, 0.21]$, was computed. $Z = 1$ means that treatment was received and $Y = 1$ implies a negative outcome. Note that the respective proportions of the PSY 101 data (Section 5) are within this range ($P(Y_{\text{PSY101}}) \approx 0.15$, $P(Z_{\text{PSY101}}) \approx 0.42$).

Ten covariates $(X_1, \ldots, X_{10})$, six binary $(X_1, X_3, X_5, X_6, X_8, X_9)$ and four continuous (X_2, X_4, X_7, X_{10}), were introduced. Four confounders $(X_1, \ldots, X_4)$ were generated, associated with both the treatment variable Z and the outcome variable Y. Three instrumental variables (X_5, X_6, X_7) were associated with just the treatment Z and three variables (X_8, X_9, X_{10}) with the outcome Y only. In the following, we explain the data generation process in detail.

The ten final covariates $(X_1, \ldots, X_{10})$ were generated in two steps. First, eight base covariates $(V_i, i \in \{1, \ldots, 8\})$ and two final covariates (X_7, X_{10}) were generated as independent standard normal distributed random variables. The remaining eight final covariates were computed by a linear combination of $(V_1, \ldots, V_8)$ to introduce correlations. Low correlations (0.2) and high correlations (0.9) were generated. After this step, all variables $(X_1, \ldots, X_{10})$ were rescaled to mean zero and standard deviation one. The correlation matrix, as well as the linear combinations, are shown in Table A.5 in the Appendix. Correlation magnitudes were reported before dichotomizing, which diminishes the correlations. For dichotomizing, we used the theoretical (0.7, 0.65, 0.6, 0.55, 0.5, 0.4)-percentiles of the standard normal distribution as cut-off points. Consequently, on average this led to prevalences of $(30\%, 35\%, 40\%, 45\%, 50\%, 60\%)$ for the binary variables $(X_1, X_3, X_5, X_6, X_8, X_9)$, respectively.

The simulated "true" propensity score $e(X) = P(Z = 1|X)$ was generated using logistic regression as a function of the four confounding and three instrumental covariates X_i.

$$e(X) = f(X_i; \beta), \qquad i = 1, \ldots, 7, \tag{6}$$

where the function f and the parameters β vary in each scenario to introduce different levels of non-linearity and non-additivity in the propensity score model. The functions f and parameters β are listed for each scenario in the Appendix (Section 8, Table A.7). We randomly sampled from the simulated "true" propensity score $e(X)$ to obtain the binary treatment variable Z. More precisely, for each data set we generated a random vector $u \overset{iid}{\sim} Uniform(0, 1)$ and let

$$Z = 1 \Leftrightarrow e(X) > u. \tag{7}$$

The binary outcome variable Y was modeled in a similar way. Logistic regression was used as a function of the seven outcome related covariates $X_i, i \in \{1, 2, 3, 4, 8, 9, 10\}$, and treatment Z to calculate the probability of Y, given X_i and Z as

$$P(Y = 1|Z, X_i) = [1 + \exp\{-(\alpha_0 + \sum \alpha_i X_i + \gamma_{LO} Z\}]^{-1}. \tag{8}$$

Values for α_i are in Table A.6 in the Appendix. The same logistic regression model (8) was used for all scenarios. The binary outcome Y was obtained analogously as in (7), again generating $u \overset{iid}{\sim} Uniform(0, 1)$.

We set the effect of treatment Z to a range of constant values $\gamma_{LO} \in \{-1.2, -1.0, -0.8, -0.6, -0.4, +0.4\}$, which we consider as realistic in the context of student success studies. As a comparison, in our data analysis in Section 5, we obtain an average treatment effect estimate of $\hat{\gamma}_{LO} \approx -0.52$. Recall that $Y = 1$ is a negative outcome. Therefore, a negative γ_{LO} represents a reduction of negative outcomes and thus a positive treatment effect.

Note that the values γ_{LO} denote the conditional log-odds ratios, relating treatment to the outcome. This is the ratio of the odds that an observation has outcome $Y = 1$ when it is treated, versus the odds that $Y = 1$ when it is not treated, given in log scale. To obtain the true marginal risk differences γ_{RD}, which we aim to estimate (Formula 4), we follow the procedure described in Austin (2010). Assuming that the entire population was untreated, the probability of the outcome $Y = 1$ was computed for each observation, respectively. Austin (2010) then denote the average of these outcome probabilities as the marginal probability of the outcome if untreated. Analogously the marginal probability of the outcome if treated was computed, by assuming that all observations receive treatment. Then the marginal risk difference, also denoted as population-average risk difference, is given by the difference between the two marginal probabilities. Averaged over the $m = 1000$ simulated data sets, we then obtained risk differences of $\gamma_{RD} = -0.103, -0.091, -0.078, -0.062, -0.044$ and $+0.055$ for $\gamma_{LO} = -1.2, -1.0, -0.8, -0.6, -0.4$, and $+0.4$, respectively.

3.2. Simulation scenarios

The following propensity score scenarios were modeled. The associations of the covariates with the propensity score are in the log-odds scale. Parameters and formulas are in Appendix A (Section 8, Table A.7):

- A: Additivity and linearity (only main effects);
- B: Mild non-linearity (one quadratic term);
- C: Moderate non-linearity (three quadratic terms);
- D: Mild non-additivity (four two-way interaction terms);
- E: Mild non-additivity and non-linearity (four two-way interaction terms and one quadratic term);
- F: Moderate non-additivity (nine two-way interaction terms);
- G: Moderate non-additivity and non-linearity (ten two-way interaction terms and three quadratic terms);
- H: Severe non-additivity and non-linearity (six two-way interaction terms, four three-way interaction terms, one quadratic term, one cubic polynomial, and one square root term).

To ensure robust and realistic scenarios, we attached importance to the simulated propensity score distributions. The simulated propensity scores meet the non-zero probability assumption (2), and overall less than 1% of the propensity score values fall below 0.01 or above 0.99. We suppose that this is a realistic assumption for the majority of educational, observational treatment effect studies. Further, we found it realistic that there is a positive probability (support) for propensity scores across the entire $(0, 1)$ range, yet neither uniformly distributed nor heavily skewed. Thus, with some variations across scenarios, our simulated propensity score distributions are right-skewed with a mode between 0.1 and 0.25. A histogram of the composed propensity scores over all simulated scenarios collectively is illustrated in the Appendix in Figure A.4.

3.3. Performance metrics

Recall that our Monte Carlo simulation study comprises three different data sizes, eight scenarios with varying associations of the covariates with treatment assignment, and six simulated treatment effect magnitudes, leading to overall 144 different setups. To get a Monte Carlo estimate and performance assessment of our proposed stacked ensembles, the following performance measures were computed based on $m = 1000$ simulated data sets for each setup, respectively.

- **Bias:** The mean difference of the average treatment effect estimates $\hat{\gamma}_{RD}$ (Formula 4) and the simulated (true) treatment effects γ_{RD}. The absolute value of the relative bias (in %) is reported by dividing through the simulated (true) treatment effect γ_{RD}, respectively.

- **MSE:** The mean squared error of the treatment effect estimates $\hat{\gamma}_{RD}$ with the true treatment effect γ_{RD}.

- **MC-SE:** The empirical Monte Carlo standard error, calculated as the standard deviation of the average treatment effect estimates $\hat{\gamma}_{RD}$ over the $m = 1000$ simulations.

- **SE:** The estimated standard error of the treatment effect estimates, averaged over the $m = 1000$ simulations. We used a robust sandwich-type variance estimator, implemented in the `survey` package in R (Lumley, 2004). Correlations are introduced in the sample through replications of subjects, based on the subject-specific IPTW weights (Formula 3). The objective of this robust sandwich-type variance estimator is to provide consistent variance estimation with dependant data (for a discussion, see Kauermann and Carroll 2001; Lunceford and Davidian 2004; Joffe et al. 2004).

- **ASAM:** The average of the standardized absolute mean differences (Formula 5), assessed in the original and in the weighted sample. For each model, the magnitude of the ASAM was computed in its weighted sample, and the average over the $m = 1000$ data sets was reported. Additionally, correlations between the ASAM and the estimated ATEs were assessed in each data set including several models.

- **Weights:** The distribution of the weights used in IPTW. Extreme weights might be of concern, as they lead to strong influence of the respective observations in the estimation of the average treatment effect and to increased standard errors in the ATE estimate (Cole and Hernán, 2008; Austin and Stuart, 2015).

The computation of the propensity scores in this comprehensive Monte Carlo simulation setup turned out to be highly CPU-intensive. To make the computation more feasible, in each of the $m = 1000$ data sets, we performed only one outer cross-validation fold k of Algorithm 1 as sketched in Figure 1. The propensity scores and the above performance measures were then assessed on a validation set D_k containing one-quarter of the data sets. More precisely, in Algorithm 1, we set $K = 4$ and $J = 2$, that means for the data sizes $n_{total} \in \{1000, 2000, \text{and } 3000\}$

we obtained validation sets D_k with size $|D_k| \in \{250, 500, \text{and } 750\}$ and ensemble training sets $|E_k| \in \{750, 1500, \text{and } 2250\}$, respectively. For consistency, the assessment of all models discussed in the following Section 4 was carried out on the exact same validation subsets D_k. As in Setoguchi et al. (2008) and Pirracchio et al. (2014), all 10 covariates were included as features in the propensity score estimation models.

3.4. Comparison models

We compared the performance of our proposed models GBM-Stack and LR-Stack to state-of-the-art models in propensity score estimation, and to each of the 12 included base learners (described in Table A.4). The most competitive models are presented in the result section. Precisely,

- Twang-GBM – using the `twang-package` (Ridgeway et al., 2014) with the same hyperparameters as used by Lee et al. (2010), recommended by McCaffrey et al. (2004), with 20,000 iterations and a shrinkage parameter of 0.0005, minimizing the mean of the Kolmogorov-Smirnov statistic;
- Superlearner (SL) – proposed by Pirracchio et al. (2014), with implementation as described by Naimi and Balzer (2018), including all 12 base learner models (Appendix, Table A.4) with maximization of the area under the ROC-curve;
- NNET – a single-hidden-layer neural network;
- AVNNET – a model average of several independently-initialized and in parallel trained neural networks to prevent overfitting and overconfident predictions;
- LR – a simple logistic regression model;
- RLR – a regularized logistic regression model, adding a penalization of high parameters to the logistic regression model to prevent overfitting.

4. Results

We emphasize that our main target of the propensity score estimation procedure is to reduce bias in the population average treatment effect estimation (ATE) obtained through Formula 4, following Lunceford and Davidian (2004). Due to reasons of space and clarity, in the following we outline a summary of the results of the various simulation setups. We refer to our Appendix for an exhaustive result listing of all individual simulation setups and performance metrics.

4.1. Performance of the ATE estimator

4.1.1. Bias of ATE estimation

Table 1 serves as an overall summary of the simulation study results by averaging the assessed performance metrics obtained for the 48 different simulation setups (eight different scenarios (A-H) and six different simulated ATEs γ_{RD}) for each of the three data sizes, respectively. The first column of Table 1 presents the resulting summarized relative bias (in %) for the different propensity score-based ATE estimators. In summary, both of our proposed stacked ensembles

Table 1: Summary of the simulation study performance metrics averaged across all eight scenarios A-H and across the six simulated ATEs (γ_{LO}) for each data size, respectively. The row "Simulated PS" presents the results obtained by incorporating the respective simulated "true" propensity scores in the ATE estimation.

n= 3000	Bias (%)	MSE	MC-SE	SE	ASAM	Max-Weights
GBM-Stack	1.37	0.0009	0.0298	0.0293	0.088	11.65
LR-Stack	1.47	0.0009	0.0301	0.0294	0.093	12.40
Superlearner	3.87	0.0008	0.0282	0.0279	0.087	11.10
AVNNET	1.97	0.0010	0.0307	0.0301	0.091	21.56
NNET	1.62	0.0012	0.0343	0.0323	0.100	33.00
RLR	3.68	0.0009	0.0287	0.0280	0.097	25.85
LR	4.23	0.0009	0.0290	0.0282	0.100	27.81
Twang-GBM	8.26	0.0008	0.0274	0.0275	0.090	13.13
Simulated PS	1.02	0.0012	0.0348	0.0325	0.088	31.92
n = 2000	Bias (%)	MSE	MC-SE	SE	ASAM	Max-Weights
GBM-Stack	1.02	0.0013	0.0364	0.0355	0.107	11.10
LR-Stack	1.16	0.0014	0.0370	0.0360	0.115	13.13
Superlearner	1.99	0.0012	0.0349	0.0341	0.104	10.07
AVNNET	1.65	0.0015	0.0382	0.0367	0.111	20.55
NNET	2.77	0.0019	0.0427	0.0395	0.125	31.10
RLR	4.00	0.0013	0.0356	0.0340	0.115	22.93
LR	4.76	0.0014	0.0362	0.0344	0.120	25.38
Twang-GBM	8.94	0.0012	0.0336	0.0334	0.104	13.06
Simulated PS	0.99	0.0018	0.0427	0.0390	0.107	27.44
n = 1000	Bias (%)	MSE	MC-SE	SE	ASAM	Max-Weights
GBM-Stack	2.17	0.0026	0.0503	0.0495	0.149	10.24
LR-Stack	2.55	0.0028	0.0524	0.0504	0.162	12.88
Superlearner	3.77	0.0023	0.0473	0.0469	0.142	8.33
AVNNET	3.28	0.0032	0.0563	0.0516	0.162	20.72
NNET	4.76	0.0039	0.0614	0.0550	0.182	28.35
RLR	4.13	0.0025	0.0490	0.0474	0.157	18.32
LR	5.48	0.0027	0.0504	0.0481	0.164	21.60
Twang-GBM	10.67	0.0022	0.0462	0.0464	0.139	10.81
Simulated PS	1.90	0.0034	0.0577	0.0527	0.147	20.50

(GBM-Stack and LR-Stack) demonstrated superior reduction of bias in the ATE estimation for all three data sizes. With an average relative bias of 1.37%, 1.02% and 2.17% for the respective data sizes $n_{total} \in \{3000, 2000, \text{and } 1000\}$, especially the GBM-Stack-based ATE estimator led to nearly unbiased ATE estimates. The performance almost reaches the top benchmark of 1.02%, 0.99%, and 1.90% relative bias, which was obtained by including the simulated "true" propensity scores (from Formula 6) in the ATE estimation (obviously not available in practice). Note that ignoring confounding by just taking the differences of outcomes of the treated and untreated group (without using propensity score weighting) would have led to significantly bi-

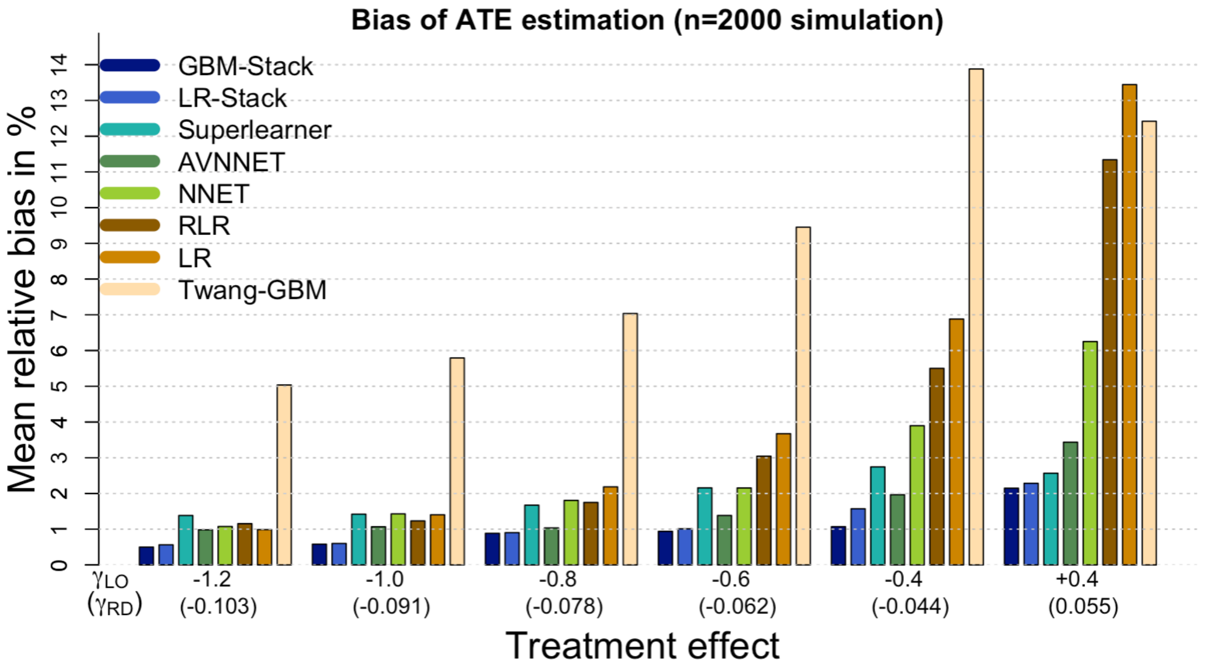

Figure 2: Comparison of relative bias (in %) resulting from the different ATE estimators for data size $n = 2000$. The illustrated bias results present the average bias over the eight scenarios (A-H) for each treatment effect $\gamma_{LO} \in \{-1.2, -1.0, -0.8, -0.6, -0.4, +0.4\}$, respectively. For data size $n \in \{3000, 1000\}$, the results are similar and illustrated in Figure A.5 and A.6 in the Appendix.

ased ATE estimates. On average across simulation setups such a "naive" approach would have led to ~34% relative bias in the ATE estimator, ranging between 5% and 80% across setups (Appendix, Tables A.32, 66, and 100), thus demonstrating the positive impact of the described stacked ensemble propensity score based IPTW approach to substantially reduce bias in the ATE estimation.

Comparing the performance of our stacked ensembles with other elaborated propensity score estimators proposed in the literature, most notably, the high bias of the Twang-GBM based estimator (Ridgeway et al., 2014) is apparent. With an overall average bias of 8.26%, 8.94%, and 10.67% for the three data sizes (Table 1), the Twang-GBM based estimator failed to remove bias in the ATE estimator to a large extent. Using Superlearner (Pirracchio et al., 2014) led to less biased estimates than the Twang-GBM model, yet overall on average still around twice the magnitude of our GBM-Stack model bias. Figure 2 provides a more detailed comparison of the resulting bias of the different propensity score-based ATE estimates. More precisely, Figure 2 presents the average relative bias (in %) obtained for each treatment effect setup $\gamma_{LO} \in \{-1.2, -1.0, -0.8, -0.6, -0.4, +0.4\}$ for data size n=2000 by averaging the bias results of scenarios A-H, respectively. For data sizes $n \in \{3000, 1000\}$ the respective figure is found in the Appendix (Figure A.5 and A.6), exhibiting a very similar pattern.

Figure 2 (as well as Figure A.5 and A.6) illustrate that both stacking models GBM-Stack and LR-Stack led to the lowest bias in the ATE estimator across all considered simulated treatment effects $\gamma_{LO} \in \{-1.2, -1.0, -0.8, -0.6, -0.4, +0.4\}$. On the contrary, the Twang-GBM model (Ridgeway et al., 2014) led to substantially biased ATE estimators, with up to ~14% relative bias for $\gamma_{LO} = -0.4$ and data size n=2000 (Figure 2), whereas GBM-Stack led to relative bias of ∼1% in the same setup. Superlearner (Pirracchio et al., 2014) led to moderate relative bias results across the simulated ATE, between two and three times as high compared to GBM-Stack, except for $\gamma_{LO} = +0.4$ where it led to only slightly higher bias. Note that especially for the larger data size $n = 3000$ (Appendix, Figure A.5) the resulting bias of Superlearner was considerably higher compared to the other models, with a maximum of ∼6% relative bias for $\gamma_{LO} = -0.4$. We also compared our models to a standard logistic regression (LR) estimator, conventionally used for propensity score estimation in the literature. The logistic regression-based estimator led to low/moderate bias for ATE estimates with larger treatment effect magnitudes ($\gamma_{LO} \in \{-1.2, -1.0, -0.8\}$). However, a large increase of relative bias can be seen for $\gamma_{LO} \in \{-0.6, -0.4, +0.4\}$ setups, leading to a maximum relative bias of ∼13.5% for $\gamma_{LO} = +0.4$ (Figure 2). We found that the regularized logistic regression version (RLR), which aims to prevent overfitting by adding a penalization of high parameters to the logistic regression model, is favorable in terms of bias reduction across the ATE setups. Regularized logistic regression led on average to around 1% less bias than the standard logistic regression model (LR). More competitive in terms of bias reduction were the neural network-based estimators NNET and AVNNET. Across the different ATE setups of data size n$\in \{2000, 1000\}$ (Figure 2 and A.6) the AVNNET-based estimator led to less bias than NNET (of around 1% on average). In these setups, AVNNET was thus most competitive to our stacked models in terms of bias reduction. NNET led to slightly less bias than AVNNET across the $n = 3000$ ATE setups. However, both neural network-based models still led to higher bias than our stacking models. Finally, comparing both of our proposed methods GBM-stack and LR-stack, a slight but consistent advantage is visible in favor of the GBM-Stack model across the ATE setups. There was only one ATE setup, $\gamma_{LO} = -0.4$ with $n = 3000$ (Figure A.5), where LR-Stack led to slightly less bias than GBM-Stack.

Note that the general increase of the presented bias results from $\gamma_{LO} = -1.2$ to $\gamma_{LO} \in \{-0.4, +0.4\}$ (Figure 2) is partially due to the fact that we present the relative bias by dividing the absolute bias magnitudes through the respective ATE magnitude $|\gamma_{LO}|$. However, the increase of bias is not proportional across the models (e.g. increase of relative bias for logistic regression is stronger). Whereas we do not have any theoretical or experimental explanation for that behavior, we assume that at some regions of the propensity score estimation (for certain observations) the propensity scores are not correctly estimated, leading to errors in the computed sample weights. In the estimation of the ATE, these weights only make a difference if the respective outcome is one (not zero). Since the marginal outcome probabilities are decreased with negative ATEs, these (incorrect) weights matter less for the strongly negative ATEs. Figure 2 (as well as Figure A.5 and A.6) demonstrate that our stacked ensembles led to the lowest relative bias throughout the different ATE setups, indicating robustness of our method with respect to changes in treatment effect magnitudes.

Since the underlying associations between the covariates and the treatment (e.g. scenarios A-H) can usually not be identified in practice, we suppose that any recommended model should perform well across a range of various scenarios. This is summarized above by the average bias across scenarios A-H (e.g. Figure 2). In the following, we briefly examine the bias results for the scenarios A-H separately. We summarized the simulation results for each scenario by averaging the relative bias results obtained for the six ATEs γ_{LO} and the three data size setups for each scenario A-H (Appendix, Table A.111), respectively. Our proposed GBM-Stack model led to the lowest relative bias in five out of eight scenarios (A, B, D, E, H), with a minimum relative bias in scenario A (0.38%) and a maximum in scenario F (2.58%). Both stacked models led to lower relative bias than the Twang-GBM model in all eight scenarios A-H, which had its maximum bias in scenario E (13.2%). Superlearner led to higher bias than both of our stacked models in all scenarios except for scenario G (1.74%), with a maximum in scenario F (5.16%). The AVNNET based model led to lower bias than our stacked models only in two out of the eight scenarios (C and F), with a maximum relative bias in scenario E (3.32%). As expected, the logistic regression-based model performed well in scenario A (0.84%), which only contains main effects in the simulated propensity score model. However, in scenarios with more complex associations between the covariates and the treatment (non-additivity or non-linearity) the logistic regression model led to strongly increased bias, with a maximum of 8.13% in scenario H. A similar trend was shown for the regularized logistic regression model, which had slightly less bias compared to the plain logistic regression model across scenarios.

4.1.2. Variability and mean squared error of ATE estimation

In addition to the main objective of our study, the bias of the ATE estimators, we were also interested in the variability of the estimates, which we assessed through the Monte Carlo standard error of the $m = 1000$ ATE estimates (MC-SE) in each simulated setup. A summary of the resulting MC-SE for our three data size setups (averaged over the eight scenarios and six ATEs) is presented in the third column of Table 1. In general, a variance-bias trade-off is visible when comparing the bias and MC-SE across models in Table 1, whereas models with lower bias tended to have more variability, expressed in higher MC-SE. The Twang-GBM model, with the most strongly biased estimates among the compared models, led to the lowest variability with an average MC-SE of 0.0274, 0.0336, and 0.0462 for the data sizes n$\in \{3000, 2000, 1000\}$. Superlearner, with moderately biased estimates, led to slightly higher MC-SE of 0.0282, 0.0349, and 0.0473 on average for the three data size setups. On the other hand, the relatively little biased neural network-based estimate (NNET) led to the highest variability with an average MC-SE of 0.0343, 0.0427, and 0.0614 for the three data sizes, respectively. Note that using the averaged neural network model (AVNNET) led to substantially improved variability compared to the single neural network (NNET) with around 10% less MC-SE on average across the simulated setups. Similarly, the regularized logistic regression model (RLR) had consistently less MC-SE than the plain logistic regression model (around 2% on average). Examining the variability of our proposed stacked ensemble model estimates, the variance-bias trade-off was less substantial. Having the lowest biased estimates across setups, our stacking models still had lower MC-SE than both neural network-based models and a similar magnitude of MC-SE compared to the moderately biased logistic regression model. Comparing both stacking models, with an average MC-SE of 0.0298, 0.0364, and 0.0503, the GBM-Stack model led to slightly less variability than the LR-Stack based estimates (0.0301, 0.0370, and 0.0524).

In general, we suppose that overall the magnitude of variability of the presented ATE estimators is reasonably small. Except for the neural network (NNET) based model, all discussed propensity score models led to lower variability in the ATE estimation (through Formula 4) than incorporating the simulated “true” propensity scores. Note that this result is in line with theoretical findings (Hirano et al., 2003) and empirical findings (Griffin et al., 2017), that well-estimated propensity scores can lead to more efficient ATE estimators by incorporating the observed treatment status.

Obviously, the MC-SE of the models is only available in a Monte Carlo simulation study with a number of i.i.d. sampled data sets for each setup. In practice, on a real-life data set, the standard errors are commonly estimated using a robust sandwich-type variance estimator (Kauermann and Carroll, 2001), as described in Section 3.3. To assess the estimates, we computed the robust sandwich-type standard errors on each of the $m = 1000$ data sets for each of the setups and report the mean of the $m = 1000$ estimates, respectively (Appendix, Tables A.26-31, 60-65, 94-99). Table 1 contains a summary of the estimated robust sandwich-type standard error estimates (SE) for the three data size setups (averaged over the eight scenarios and six ATEs), indicating lower magnitudes of estimated SE than the reported MC-SE. Considering the ratio (MC-SE/SE) across the simulation setups (Appendix, Table A.112), we found that the robust sandwich-type standard errors consistently underestimated the MC-SE for most of the models, with around 5-10% higher MC-SE for the neural network-based estimates, around 3-4% for the logistic regression-based estimates, and around 1-2% for GBM-Stack and Superlearner (SL). For the Twang-GBM estimates, the estimated SE had a similar magnitude compared to the MC-SE. Note that this is a quality test of the robust sandwich-type standard error estimator to accurately estimate the variability of the different ATE estimators, rather than a quality of the ATE estimators themselves. We suppose that underestimating the standard errors by 1-2%, as occurred for the GBM-Stack or Superlearner model, is not overly severe, but since it was observed consistently throughout the setups, it might evoke the necessity of a more conservative standard error estimator.

Finally, the variance-bias trade-off as described above is recognizable when comparing the mean squared errors (MSE) of the ATE estimates. Models with a low bias tended to have a higher variance, resulting in a slightly higher MSE. Table 1 shows that overall the MSE of the Twang-GBM model, Superlearner, and our proposed GBM-Stack model is very similar. With an MSE of 0.0008, 0.0012, and 0.0022, Twang-GBM had the lowest MSE values on average for the three data size setups, closely followed by the Superlearner (0.0008, 0.0012, and 0.0023) and the GBM-Stack model (0.0009, 0.0013, and 0.0026). The neural network-based models led to the highest MSE values (0.0012, 0.0019, 0.0039 for NNET), whereas the MSE of the AVNNET model was around 20-25% lower than the MSE of the single NNET model, due to its improved variability. Similarly, the regularized logistic regression model (RLR) had slightly less MSE than the plain logistic regression model (0.009, 0.0014, 0.0027). Due to overall superior bias and MC-SE results, the GBM-Stack model also had lower MSE than the LR-Stack model.

4.2. Weights assessment

Extreme weights can affect the ATE estimate significantly in IPTW (Rubin, 2001). Therefore, weight distributions with low spread and without large outliers are desirable, such that single observations do not disproportionately affect the ATE estimation. Note that all weights are greater than 1 by definition (Formula 3). The weight distributions of the presented models had an average between 1.7 and 2.3 throughout the setups, with lower quartiles close to 1.2 and upper quartiles between 1.8 and 2.2 (Appendix, Table A.39-41,73-75,107-109). High weights occur when estimated propensity scores are close to zero or one and do not match with the true class labels. In Table 1, the averaged maximum weights over $m = 1000$ simulations are shown for the three data size setups (averaged over scenarios and ATEs). On average, the Superlearner and the GBM-Stack model led to the lowest maximum weights (between 8 and 12), directly followed by the LR-Stack and the Twang-GBM model (between 10 and 13). We suppose that these maximum weights are reasonably small, such that single observations do not overly affect the point estimate. On the other side, the neural network-based ATE estimator had substantially higher maximum weights (around 30), corresponding to high variability in the ATE estimation. Note that the relatively high averaged maximum weights of the logistic regression-based models (between 21 and 27) are partially explained by extreme weights in scenario D, with maximum weights at around 50 (Appendix, Table A.38, 72, 106). In said scenario, the logistic regression-based model also had its highest variability in the ATE estimation, with a maximum MC-SE of 0.0812 (in the $n_{total} = 1000$, $\gamma_{LO} = 0.4$ setup; Appendix, Table A.93), even falling behind the neural network-based models. The prevention from overfitting, as specifically aimed by the averaged neural network (AVNNET) and the regularized logistic regression (RLR) model, led to a prevention of overconfident predictions and thus to smaller weight outliers compared to their respective plain models (NNET and LR). Recall that this was accompanied by less variability of the ATE estimates of AVNNET and RLR compared to NNET and LR, respectively.

The comparably high maximum weights for the simulated "true" propensity score resulted from our relatively small simulated probabilities of treatment assignment (Appendix, Figure A. 4), which we found realistic in educational studies, e.g. Guarcello et al. (2017). Our stacked ensembles, therefore, show a desired ability by predicting propensity scores with less confidence towards the edges (0,1), leading to low maximum weights and low variability in the ATE estimate, but still leading to unbiased point estimates. Note that for our stacking models we have not observed any cases with extreme weights, accompanied by high ATE estimation errors. Nevertheless, we suppose that particular caution and further analysis is appropriate if such weights might occur on a real-life data application.

4.3. Covariate balance – ASAM assessment

As a predominantly used balance measure, we assessed the averaged standardized absolute mean differences (ASAM), in the IPTW weighted data of the different simulation setups, with a summary presented in Table 1. The ASAM in the unweighted data was around 0.27, averaged across the data setups. In general, all presented methods led to considerable balancing of the data in the IPTW weighted data sets. Averaged across all setups, Superlearner (0.099), GBM-Stack (0.100), and Twang-GBM (0.100) led to the lowest ASAM magnitudes, very similar to the value obtained by using the simulated "true" propensity score (0.100). Slightly higher ASAM was obtained by the LR-Stack model (0.108), the neural network-based models NNET (0.116), and AVNNET (0.105), as well as the logistic regression-based models LR (0.113) and RLR (0.109). Note that the ASAM values in our larger data size setups were generally smaller due to reduced sampling error. The magnitude of the weighted ASAM results in our scenarios is comparable to results in similar simulation studies, e.g. (Lee et al., 2010; Pirracchio and Carone, 2018). In addition to the ASAM, we also computed the average of the standardized absolute mean differences including the four confounding covariates only (ASAM_{conf}), which led to very similar results, indicating that our models led to well-balanced confounders. Details are presented in the Appendix, Section 11.1.

4.4. Relation of ASAM and ATE estimates

In addition to the magnitude of ASAM in the weighted data, we were interested in the relation of the ASAM results with the performance of our models to estimate the ATE. In particular, we were interested if the ASAM can be employed as a selection measure to indicate the top-performing propensity score estimation models for a given data set in a specific scenario.

Recall that the unweighted data had an average ASAM of 0.27 leading to an unweighted "naive" ATE estimate with ~34% bias on average, whereas the average ASAM of the IPTW weighted data was between 0.10 and 0.12 with an average of around 1% to 10% relative bias in the ATE estimates across the models. Thus, comparing the non-weighted and the IPTW weighted data, the substantial decrease of ASAM (Section 4.3) was generally accompanied by a reduction of bias in the ATE estimates for all presented models (Section 4.1.1). As a comparison, using a very poor propensity score estimation model such as bagged CART (as described in Table A.4 in the Appendix) led to highly imbalanced covariates with an average ASAM of around 0.30 and an average of ~25% relative bias. The failure of removing confounding associated with high bias could thus be well indicated by the high magnitude of ASAM and distinguished from our presented "well" performing models. Considering the scenarios individually, for instance, the previously in Section 4.2 mentioned poor performance (in terms of bias and MC-SE) of the logistic regression-based model estimate in scenario D was also associated with higher ASAM values (around 0.170), substantially higher than the average ASAM of logistic regression across scenarios (around 0.113) (Table A.36, 70, 104).

In these preceding simulation examples, ASAM successfully measured imbalance and indicated confounding in the data, connected to poor ATE estimation. This serves as a motivation to use the ASAM as a balance check assessing the fit of the propensity score estimation models as proposed in the literature (Austin and Stuart, 2015). Having said this, we were further interested if the ASAM also contains meaningful information about the quality of the ATE estimate given a specific data set application and a set of different ATE estimators. More precisely, we were interested if it is a reasonable approach in practice to fit various propensity score models on a given data set and select the model with the lowest ASAM, as one might be tempted to do, given the previous results. Note that by directly targeting the optimization of ASAM, Ridgeway et al. (2014) carry out a similar strategy, though only considering the ASAM in the Twang-GBM model fit and not comparing the ASAM of a set of several different estimation models. To investigate our question, in each scenario and in each of the $m = 1000$ data sets, we used the absolute ATE estimation error and the ASAM (or ASAM$_{conf}$) from our eight examined learners and calculated the Pearson correlations over these eight data points. We then calculated the average of the correlations over the $m = 1000$ data sets in every scenario. Recall that in this case a correlation of 1 means that a high absolute ATE estimation error is directly connected to a high ASAM in each data set. To our surprise, we did not find any correlations in this examination. Using the ASAM, the average correlation across setups was 0.06, with an overall maximum of 0.20 in scenario D in the $n_{total} = 1000$ and $\gamma_{LO} = 0.4$ setup, containing the poor performance of the logistic regression-based model as previously mentioned. Using Spearman correlations, or, the ASAM$_{conf}$ instead of the ASAM, led to even lower correlations with a mean of 0.05 across the setups. These results suggest that, given a specific data set in a specific scenario, the ASAM does not contain helpful information about the performance of the models examined in our simulations (summarized in Table 1). Thus, a selection of a specific model based on the ASAM magnitude obtained on a given data application might be misleading. Note that adding two very poor performing models KNN and BAG-CART with high ASAM and high ATE estimation error led to considerably higher correlations at around 0.5 averaged across all setups. However, we suppose that in practice these two models would not be considered anyway due to their overall poor performance. We refer to the Appendix for a listing of the computed correlations (Table A.113 - 126). In addition, in our simulations, we could not find any indication of a critical magnitude of standardized absolute mean differences that would generally imply concerning confounding and bias in the point estimate (details provided in the Appendix, Section 11.2).

In summary, our simulations suggest that the ASAM is not a suitable measure for model comparison and selection, given a specific data application and the set of ATE estimation models examined in our simulations. With the lack of such a measure, we cannot indicate a certain expert model which works well on a given data setup. This strengthens our motivation to use an estimation model that works well on a broad range of setups, such as our proposed GBM-Stack model. Note that we do not generally challenge the usage of the ASAM as a balance measure. We still suppose that computing and comparing the ASAM of the unweighted and model weighted data, as proposed in the literature (Austin and Stuart, 2015), is a reasonable course of action in practice, mainly for two reasons. Firstly, the reduction of bias in the ATE estimate from the unweighted to the IPTW model weighted data was accompanied by a reduction of ASAM throughout setups and models, as specified in the first paragraph of this section. We suppose that an unbiased model should demonstrate similar behavior in practice. Secondly, our proposed stacked models performed very well across the examined data setups and no erroneous or poor model fits were observed, accompanied by high ASAM. We cannot rule out that such poor model fits never occur in practice, for whatever reasons. The assessment of ASAM thus serves as a sanity check, whereas excessive ASAM values, as an indicator of imbalance in the weighted data, should raise concerns. We suppose that some expert knowledge in the field of application is necessary to assess which magnitudes of ASAM, or more specifically which standardized absolute mean differences for the individual covariates might indicate concerning confounding for the specific application.

4.5. Summary of main simulation results

- Our proposed stacked ensembles LR-Stack and GBM-Stack demonstrate superior bias reduction in ATE estimation using IPTW compared to state-of-the-art propensity score estimation models on a broad range of realistic educational data setups;
- GBM-Stack leads to lower bias and lower variability in ATE estimation than LR-Stack. We thus suggest it as the novel default method for propensity scores when using IPTW;
- Our stacked ensembles lead to desirably low maximum IPTW weights, reducing the risk of excessive influence of single observations in the ATE estimation;
- Averaged standardized mean differences (ASAM) fail as a measure for propensity score model selection in our simulated data;
- The estimated robust standard errors (SE) consistently underestimate the empirical Monte Carlo standard errors, evoking the need for a more robust standard error estimator.

5. Data analysis - PSY 101

We apply the best-performing model from our simulation study, namely GBM-Stack, to a real-world data set. We investigate the effect of a Supplemental Instruction (SI) program in an introductory psychology course at San Diego State University. More precisely, we are interested in the population average treatment effect (ATE) of the SI program on student success. Since attendance at the SI program is voluntary, selection bias leads to differing covariate distributions in the treatment and control groups. To get an unbiased estimate of the ATE, we employ inverse probability of treatment weighting (IPTW) as described in Section 2.1 (Formula 4), based on our proposed GBM-Stack estimated propensity scores. The purpose of the analysis in this section is twofold. It firstly serves as an illustration of our proposed stacking method on real-world educational data. Secondly, the analysis itself contributes to the educational literature, since the population average treatment of the SI course, implemented at San Diego State University, has not yet been assessed before. In Section 6, we discuss the difference to a propensity score matching study on previous SI data by Guarcello et al. (2017), which assesses the average treatment effect of the actually treated group (ATT), with an emphasis that the ATT and ATE are two very distinct concepts and statistics.

5.1. PSY 101 Supplemental Instruction (SI)

The data was collected from two sections in an introductory psychology course (PSY 101) in Fall 2015, 2016, and 2017, respectively. The courses were taught by the same instructor. In addition, the two sections had the same syllabus, and an identical structure in terms of textbook, homework, and exams. The sections were given in a hybrid format, where the students had one day of a face-to-face lecture, and a live online lecture the other day. Students who were unable to attend the online sessions were able to access the lecture at their convenience since the lectures were recorded.

This introductory course is required for psychology majors and it is a prerequisite in social and behavioral sciences for other majors. It is indeed a course in which many freshmen and sophomores will enroll to fulfill their general education requirements. This course has a high DFW rate (grades D, F or withdraw W), therefore it has been identified by the California State University (CSU) system as a bottleneck. This increases the chances of not having enough seats available for incoming freshmen or transfer students due to many students having to repeat the course. Consequently, such bottleneck courses have impacted the university by slowing down student progress towards graduation.

The university provides SI to address a significant challenge, bottleneck courses. SI was created at the University of Missouri Kansas City (UMKC) in 1973. SI started at San Diego State University (SDSU) in Fall 2015 following the model created at UMKC (UMKC, 2018). SI leaders are undergraduate students that successfully completed the given course. The SI leaders must have an overall grade point average of 3.0, earned at least a B+ when they took the course, and ought to be recommended by the professor of the course. The SI leaders are trained in state-of-the-art methods for active learning environments and the SI peer-assisted learning infrastructure. The SI leaders attend class to serve as model students for the enrolled students and lead scheduled weekly 90-minute SI sessions available to all students enrolled in PSY 101. Student participation in these SI sections is voluntary and non-credit bearing.

5.2. Data description

The PSY 101 data set contains 2,173 observations and 33 covariates. In total, 911 students attended the SI program at least once which corresponds to a proportion of 0.42 of the data. The response variable used for this study was a binary pass/fail outcome. The cut-off to pass the course was a C- or better; and D, F, or withdraw (W) was considered as failed. This so-called DFW outcome is a standard measure at CSU to identify student success. Note that student success can also be determined more generally by the learning progress or knowledge of a student at a certain time – we refer to Pardos et al. (2012) for more details. There is a 15% proportion of students earning a D, F, or W grade. There are 27 categorical variables and 6 continuous variables for our explanatory covariates which we consider as predictive for both SI attendance and course success. Table 2 contains a short description of each covariate, and more detailed information is provided in Table B.135 in the Appendix.

The data set was almost complete, 97% of the observations did not have missing values. Just four variables, SAT Comp Conv, EOT Term Units Enroll, EOT Term Units Earn, and Incoming GPA, had missing values with a proportion of 2.2%, 1.8%, 1,8%, and 0.1%, respectively. The values were imputed for each split, as described in Algorithm 1, using k-nearest neighbour imputing in the `caret` package (Kuhn, 2008).

5.3. Propensity score and ATE estimation

We used Formula 4 (Section 2.1) to estimate the ATE of SI attendance with respect to student performance in the PSY 101 course. To estimate the propensity scores, we set up GBM-Stack as described in Algorithm 1 (Section 2.2) using GBM as a meta learner and all 12 models (described in the Appendix, Table A.4) as base learners. We performed $K = 3$ fold outer cross-validation for the ensemble evaluation (step 2-5 and 12-18 in Algorithm 1) and a $J = 2$ fold inner cross-validation to obtain the level-1-data (step 6-11 in Algorithm 1). The base and meta learner training and tuning parameter selection was implemented such as in our simulation study, described in Section 2.3.

We used bootstrapping to estimate the standard errors of the ATE estimates since the robust sandwich-type standard error estimates (described in Section 3.3) underestimated the empirical Monte Carlo standard errors throughout the simulation studies (Section 4). To reduce computational cost, we bootstrapped the estimated propensity scores after the model fit and not the actual data set before model fit. $B = 10,000$ bootstrap samples were created and used to estimate 95% confidence intervals of our point estimates.

5.4. Results

5.4.1. Covariate balance and weights

We assessed the balance of the PSY 101 data before and after weighting following the recommendations by Austin and Stuart (2015). Our evaluations indicated moderate covariate imbalance in the unweighted data, which could be reduced by IPTW, based on our GBM-Stack estimated propensity scores. There were eight covariates with standardized mean difference (smd) greater than 0.1 in the unweighted data, which was reduced to only one covariate in the weighted data. In the unweighted data, the average of the standardized absolute mean differences (ASAM) over the 33 covariates was 0.066 and the mean Kolmogorov-Smirnov statistics (K-S) across all covariates was 0.0154, with a maximum at 0.0696. In the IPTW weighted data, the overall ASAM (including all 33 covariates) was 0.040, and an average K-S statistic of 0.0104 was obtained, with a maximum of 0.0483. For the purpose of our study, we were confident about the balance achieved in the data after weighting. A detailed balance assessment for each covariate is presented in Appendix B, Section 13.2.

The distributions of the weights (computed through Formula 3, Section 2.1) were centered between 1 and 2, with maxima weights below 10 in both groups. The relatively low weights show that no observations overly impact the treatment effect estimation due to large weights.

Table 2: Variable description for the PSY 101 course data.

Variable	Description
Outcome Variable	
Course success	Binary pass/fail (at C- cutoff) based on PSY 101 final grade.
Treatment Variable	
SI	Supplemental Instruction course attendance.
Categorical Covariates	
Course number	Specification of the course taken by the student.
Period	The period when the student took the course.
Entry Term	Year of first attendance in any term at California State University.
URM	Underrepresented minority, self-reported by student at application.
Gender	Self-identified gender.
Parent 1	Highest education (HE) of first parent.
Parent 2	HE of second parent. Measure of family experience in HE.
Military	Duty in U.S. military in past or dependent of active member.
In Service Area	Graduated from a school that is inside SDSU's service area.
County Name	Student's Institution of Origin County
HS Grad Year	Year student graduated from High School.
AP Calculus	Advanced placement examination Calculus.
AP Statistics	Advanced placement examination Statistics.
AP Chemistry	Advanced placement examination Chemistry.
AP Biology	Advanced placement examination Biology.
AP English	Advanced placement examination English Language
AP English	Advanced placement examination English Literature.
Fall Both	Proficient in Math and English by beginning of first Fall semester.
HS Math	Proficient in Math by end of High School.
HS English	Proficient in English by end of High School.
Compact	Indicator for Compact Scholars program participants.
EOP	Indicator for Education Opportunity Program participants.
FAST	Indicator for FAST program participants.
Res Learning Com	Indicates Residential Learning Community program participants.
Term 1 SIMS College	Indicates a student's College assigned in term 1.
Term1 Pre Major Status	Indicates whether a student is a pre-major.
Term1 Housing	Student resided in campus residence hall in term indicated.
Continuous Covariates	
Age	Student's age at entry.
SAT Comp Conv	Combined SAT Verbal and Math.
Incoming GPA	Official GPA for admission to any California State University.
Incoming Units	Transferable units earned at universities other than CSU campus.
EOT Term Units Enroll	Number of units enrolled at indicated term.
EOT Term Units Earn	Number of units earned at indicated term.

5.4.2. Treatment effect estimation

We obtained an ATE estimate of -0.066 (marginal risk difference) with a 95% bootstrap confidence interval (CI) of (-0.097, -0.034). Recall that the outcome $Y = 1$ was set as a DFW grade in the PSY 101 course and therefore a negative outcome. This said, our results show that there is a positive effect of SI attendance regarding course success.

A summary of the SI ATE estimation results is shown in Table 3. Figure 3 illustrates the ATE estimates obtained by using $B = 10,000$ bootstrap samples of our GBM-Stack estimated propensity scores. The blue circle presents the ATE estimate using the propensity scores without bootstrapping. The distribution of the bootstrap estimated ATEs is bell-shaped and centered at the ATE estimate of the actual propensity score set. We therefore constructed bootstrap confidence intervals at the 2.5% and 97.5% quantiles of the ATE distribution, leading to a 95% bootstrap CI of (-0.097, -0.034). The bootstrap standard error of the marginal risk difference ATE estimate had a magnitude of 0.0160, whereas applying the robust-sandwich type standard error estimator led to a magnitude of 0.0157. On a log-odds scale, an ATE estimate of -0.523 was obtained with a bootstrap standard error of 0.132, whereas the robust sandwich-type standard error estimator led to a magnitude of 0.130. The bootstrap standard errors were thus slightly more conservative in this analysis.

Summary Statistics	SI results
ATE (risk difference)	−0.066
ATE (odds ratio)	0.593
ATE (logOR)	−0.523
Bootstrap SE	0.016
Bootstrap SE (OR)	0.079
Bootstrap SE (logOR)	0.133
ASAM (weighted)	0.040
Mean ks-statistics (weighted)	0.010

Table 3: Summary of the population average treatment effect estimation of SI attendance on PSY 101 course success, using IPTW weights based on GBM-Stack estimated propensity scores. LogOR denotes values presented in log-odds scale.

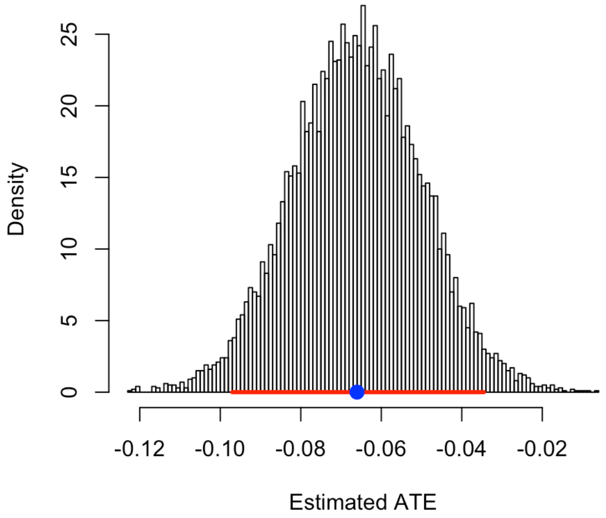

Figure 3: Histogram of estimated population-averaged treatment effects (ATE), precisely the marginal risk difference, of the SI program using $B = 10,000$ bootstrapped propensity score sets. The blue point illustrates the ATE estimate on the actual data set and the red line the 95% bootstrap CI of the point estimate.

Using this estimator, we have evidence that moving the whole PSY 101 student population to SI attendance would on average lead to 1.69 times higher odds to pass the PSY 101 class compared to not offering SI, with a 95% CI of (1.31, 2.20). We suppose that our statistically principled assessment of the effectiveness of SI on student success could have substantial implications towards the funding and extension of the program to other courses and universities.

6. Conclusion

In this paper, we present a novel stacked generalization approach for propensity score estimation used for inverse probability of treatment weighting (IPTW) to reduce bias in the estimation of population average treatment effects (ATE) in observational studies. We perform a very comprehensive Monte Carlo simulation study covering a large range of various data setups that we consider as realistic for educational, observational data. Our proposed stacked ensembles, GBM-Stack and LR-Stack, demonstrated superior bias reduction in the ATE estimation throughout the simulations, compared to currently considered models in the propensity score literature such as neural network-based models, logistic regression-based models, Superlearner, and Twang-GBM (Setoguchi et al., 2008; Lee et al., 2010; Westreich et al., 2010; Ridgeway et al., 2014; Pirracchio et al., 2014; Pirracchio and Carone, 2018), thus moving forward the best practice in propensity score estimation using IPTW. In general, most of the examined models in our simulations demonstrated well-balanced covariates and a large reduction of ATE estimation bias in the IPTW weighted data, compared to the non-weighted data – hence underlining the great value of propensity score methods on observational, educational data at large.

Our proposed stacked models led to almost unbiased ATE estimates, close to the values technically obtained by incorporating the simulated "true" propensity scores into the IPTW model, which obviously are not given in practice. It is notable that the simulated "true" propensity scores led to higher variability in the ATE estimates than our stacked ensembles. With this finding, we are in line with previous theoretical results (Hirano et al., 2003) and empirical findings (Griffin et al., 2017), suggesting that weighting on the inverse of estimated propensity scores, rather than on the "true" simulation propensity scores, can lead to a more efficient estimation of the ATE. Noise in the actual data sample is also considered through the empirical propensity score estimate. The original propensity score theory (Rosenbaum and Rubin, 1983) describes the propensity score as the coarsest balancing function, implying that more efficient balancing functions exist. Using the simulated "true" propensity scores in IPTW led to relatively large maximum weights, resulting from the low simulated propensity scores we chose to mimic educational data sets in practice. Concerning this matter, our stacking models demonstrated another desired property by having less confidence in propensity score estimates close to the edges, zero, and one. This led to low maximum weights across the scenarios.

In general, a variance-bias trade-off was visible throughout the simulations. Twang-GBM, which had the highest bias in the ATE estimation across the simulations, led to the lowest variability in the estimates. Since Twang-GBM failed to remove bias by a large proportion, thus missing the primary goal of having unbiased ATE estimates, we would not consider Twang-GBM as a good-performing model in our simulation study. Superlearner led to moderately-low biased ATE estimates, yet consistently higher than the bias of our stacked models throughout the simulation setups. Due to the variance-bias trade-off, Superlearner had lower variability in the ATE estimates and similar mean squared error compared to our proposed stacked ensembles. In this case, we suppose that unbiasedness along with accurate standard error estimates is desirable in drawing causal conclusions about treatment effects, thus favoring our proposed stacked ensemble models. In our simulations, we consider Superlearner as most competitive to our proposed models however.

The sandwich-type variance estimator (Kauermann and Carroll, 2001), commonly used to estimate standard errors in propensity score IPTW studies, slightly underestimated the variability of the ATE estimates throughout the setups. In our data analysis (Section 5), we thus used bootstrap standard errors, which were slightly more conservative in our case. We are in accordance with previous studies (Pirracchio et al., 2014), encouraging further research to obtain more accurate standard error estimates in propensity score IPTW setups.

Comparing the averaged neural network model with the single neural network, and the regularized logistic regression model with the plain logistic regression, the former overfitting prevention models led to better performance in the ATE estimation, respectively. Especially in scenarios with complex associations between the covariates and propensity scores, the logistic regression-based models led to unreliable, highly biased ATE estimates. As expected, in setups with additivity and linearity (scenario A) logistic regression led to almost unbiased ATE estimates though – showing that, if correctly specified, logistic regression is a well-performing propensity score estimation model. To explain the superior performance of our proposed stacked ensembles, the performance of the simple logistic regression model might provide some intuition. In the literature, many researchers were striving to include the correct interaction and terms of higher moments to obtain correctly fitted logistic regression models for propensity scores. However, the associations between covariates and treatment selection are usually not known in practice. Loosely speaking, by using the stacking approach this investigation is done automatically through the diverse base learner set and flexible modeling approaches therein. Thus, instead of modeling the complexity of the covariate-treatment selection associations, the logistic regression meta learner only has to learn the errors of the base learner models, which eventually led to reduced bias in the ATE estimation across the setups. Our simulations suggest that using GBM as a meta learner (GBM-Stack) even outperformed the logistic regression-based stacked ensemble (LR-Stack) in terms of bias and variability. The motivation of applying a model such as GBM-Stack, which works well on a large range of data setups, is reinforced by the next finding in our simulation study.

A primary objective of propensity score methods is to balance systematic differences in the covariates between treated and untreated observations to reduce confounding caused by selection bias. The most commonly used measure to assess balance between covariates is the standardized mean difference between covariates, recommended by Austin and Stuart (2015). In recent simulation studies (Setoguchi et al., 2008; Lee et al., 2010; Pirracchio et al., 2014; Pirracchio and Carone, 2018), the average of the standardized absolute mean differences (ASAM) is presented as a measure to assess covariate balance of the proposed algorithms. Overall, our proposed stacked ensembles showed remarkable reduction of ASAM, indicating well-balanced covariates. Investigating the relation between ASAM and ATE estimation, a decrease of ASAM in the IPTW weighted data compared to the non-weighted data was generally coherent with a decrease of bias in the ATE estimation across the examined models, confirming findings in (Franklin et al., 2014; Stuart et al., 2013; Ali et al., 2014). Given the general relation between ASAM magnitude and ATE estimation bias, we were interested if the ASAM could also serve as a measure for propensity model selection for a specific data application. However, there were no correlations between the ASAM and the ATE estimation error comparing the estimates of the examined models in our simulation study throughout the data setups. Consequently, selecting a model due to superior ASAM on a specific data set does not necessarily lead to lower error; in fact, such a choice might lead to a higher estimation error. We suppose that finding a measure that indicates a best-performing propensity score model for a specific data application is a very valuable topic for ongoing research, even having potential implications on the respective model fitting. However, the lack of such a measure evokes the need of a propensity score model that works well on a broad range of scenarios, strengthening our suggestion for the use of GBM-Stack as a default propensity score estimation model with almost unbiased ATE estimation performance throughout our simulations.

The data analysis in Section 5 serves as a demonstration of our proposed GBM-Stack model on a real-life data set, and contributes to the educational literature by providing evidence for a positive population average treatment effect of Supplemental Instruction (SI) programs (UMKC, 2018) on student success, specifically for the introductory psychology course PSY 101 at SDSU under consideration. We implemented the GBM-Stack model with the same configurations as proposed in our simulation study to estimate the propensity scores, incorporating twelve models in the base learner set (Appendix, Table A.4). This led to increased balance in the IPTW weighted data, measured by reduced standardized mean differences, Kolmogorov-Smirnov statistics, and mean differences of higher-order terms, compared to the non-weighted data. Our analysis suggests that moving the whole PSY 101 student population to attending SI would lead to 1.69 higher odds to pass the course, compared to not offering SI, with a 95% bootstrap CI of (1.31, 2.20).

We note that there has been a preceding propensity score study by Guarcello et al. (2017) on assessing the effect of SI on student success in the PSY 101 course. Guarcello et al. (2017) found that the odds of passing the course for students who attended SI were 2.2 times higher than those who did not attend any SI Sessions. We emphasize that there are several methodical differences to our analysis. The current study includes additional cohorts of more recent student observations. Further, Guarcello et al. (2017) use a cut-off at grade C, and we distinguished between a DFW and C- or better (fail\pass). Most importantly, however, Guarcello et al. (2017) use a propensity score matching approach to estimate the SI treatment effect of the actually treated group (ATT), and we estimate the population average treatment effect (ATE).

At this point, we want to highlight that the ATE and the ATT are two generally distinct estimates since individual treatment effects can be systematically different for observations (students) in the treated and non-treated groups. The underlying covariates might not only affect treatment selection, but also the effect of the treatment. As an example, well-performing students might not attend an extra tuition class if they already know the provided material, and attending the class (treatment) would not have a positive effect on them. Then, the average treatment effect might be strongly positive in the actually treated group, but negligible when considering the whole population. The benefit of the intervention would thus strongly depend on the implementation, which might be crucial in educational policymaking, such as budgeting, advertising or even obligating a treatment for the entire student population, or conceivably only for sub-groups of the population. Principally, propensity scores could also be used to estimate unbiased treatment effects for sub-populations in the data, e.g., based on categorical covariates. For instance, some universities offer voluntary scientific writing in English courses. One could, for example, estimate the ATE and ATT of those courses for international students – a proportionally high ATE might for instance justify compulsory scientific writing training for international students to prevent later bottlenecks and to increase student success. On the other side, given a comparably low ATE, making the course compulsory (for international students) might misspend resources.

Note that these are some general examples pointing out the differences between ATE and ATT with potential implications in practice. The positive ATE of SI in our study, though a bit lower than the ATT obtained by Guarcello et al. (2017), indicates that the treatment as a policy appears robust in its implementation. Estimating SI treatment effects on PSY 101 subpopulations is not directly a purpose of this paper, but an interesting application for future work. There are further extensions of the SI treatment effect analysis considered for future work. The SI treatment effect estimation could for example be compartmentalized by estimating the treatment effect conditional on the frequency of attendance, which could lead to further recommendations of how intensively the SI course should be used. This could be approached by withdrawing the assumption of a binary treatment (attendance\non-attendance), following Imai and Van Dyk (2004).

7. Limitations and Future Work

In general, there are several promising extensions of our work to consider for future research. In this paragraph, we briefly summarize additional investigations and findings in our simulation study with potential implications for further studies, as well as limitations of our work. As for any simulation study, our paper has some limitations. Even though we performed a very comprehensive simulation study, striving to imitate a broad range of real-life educational data, we cannot cover every possible data case. Since the interpretation of Monte Carlo simulation results is somehow limited to the study regime, further simulation studies might expand the generalization of our results. Besides, there are also other approaches in the statistical literature that target covariate balance and could be compared to our proposed methods in further research, for instance, entropy balancing proposed by Hainmueller (2012). Due to high computational cost, we did not compare our results to the balancing Superlearner approach (Pirracchio and Carone, 2018). However, we implemented and evaluated the Superlearner as described in Naimi and Balzer (2018) without incorporating balance measures in the optimization step. In Pirracchio and Carone (2018) the performance difference of the two Superlearner approaches was not significant. Further, in order to have a fair comparison, we included our twelve base learners in the Superlearner computation, rather than the models used in Pirracchio et al. (2014). We did not compare our ensembles to CBPS, the covariate balancing propensity score, introduced by Imai and Ratkovic (2014). However, Pirracchio and Carone (2018) already showed that both Superlearner models outperformed CBPS in their simulation study.

Note that, due to the modularity of our stacking approach, there are several opportunities for configurations that might further improve our proposed method. If further minimization regarding a certain (balance) measure is desired, one can easily increase the number of cross-folds and expand the grid for the tuning parameter selection for each base and meta learner. Further, additional outer and inner cross-validation loops in Algorithm 1 can be performed, especially on smaller data sets. Due to computational expense, we did not perform principal component analysis to reduce correlations in the level-1-data, a step that could lead to further improvements of the stacked ensembles.

To modify the level-1-data, various approaches based on different diversity and performance measures have been proposed for base learner selection (Sesmero et al., 2015; Alpaydin, 2014). We obtained some promising preliminary results by performing a pruning approach (Alpaydin, 2014, chap. 17) in part of our simulations while removing highly correlated predictors with weak performance, in terms of high log-loss (Appendix, Formula 10), from our initial set of twelve base learners. An exhaustive, iterative pruning approach on the base learner set will be considered in further research. This includes finding a reliable measure that indicates the best combination of base and meta learner.

We investigated the relationship between the propensity score estimation accuracy and ATE estimation performance of our examined models. We computed the mean squared error of the estimated propensity scores with the simulated propensity scores, as well as the Kolmogorov-Smirnov test statistics between the estimated and simulated propensity score distributions (Appendix, Tables A.127-134). However, there was no clear relation between performance w.r.t. mean squared error and Kolmogorov-Smirnoff statistics of the estimated propensity scores and the performance of the models in the ATE estimation. We suppose that finding properties in the relation of the simulated "true" propensity scores and the estimated propensity scores that allow conclusions about the performance of the estimated propensity scores in the ATE estimation is of high interest for further research. We assume that, for instance, propensity score regions close to 0 and 1 might be more important to assess, since poor prediction in these regions might lead to weights that strongly affect the ATE estimate.

Log-loss might be a useful metric for propensity score estimation since it has been shown to be advantageous in obtaining well-calibrated class probability estimates (Zadrozny and Elkan, 2001), and further highly penalizes erroneous confident propensity score predictions. Others indicate that a good propensity score model fit does not necessarily lead to good performance in ATE bias reduction (Westreich et al., 2011; Parast et al., 2017; Griffin et al., 2017). We suppose that more research has to be done to indicate properties that define a set of well-estimated propensity scores, with the ultimate goal to obtain accurate, especially unbiased ATE estimation.

Note that the modularity of our model also holds for the choice of the meta learner. In addition to logistic regression and GBM, we investigated the performance of AVNNET used as a meta learner on a subset of our simulations. The obtained AVNNET-stack results also indicated favorable bias reduction in the ATE estimation, suggesting that stacking, in general, is beneficial in propensity score IPTW ATE studies. However, AVNNET is computationally expensive, and our preliminary AVNNET-Stack results were not competitive to LR-Sack and GBM-Stack, we thus did not consider it further in our simulations. In principle, any probabilistic classification method could be examined as meta learner in further work.

We also investigated a slightly modified version of Algorithm 1, namely stacking+, as discussed by Torres-Sospedra et al. (2006; Sesmero et al. (2015). In stacking+ the level-0-data (the actual covariate data) is added to the level-1-data and the meta learner is trained on the combined set of level-1-data and level-0-data. Comparing the ATE estimation of stacking with stacking+, GBM-Stack and GBM-Stack+ led to very similar results and LR-Stack outperformed LR-Stack+ with respect to both bias and variability. Adding the higher computational expense of stacking+, we recommend the stacking approach as described in Algorithm 1.

Given the estimated propensity scores, there are additional ATE estimation methods proposed in the literature. Austin (2011) adopt a slightly different population average treatment estimation using IPTW (as shown in Formula 11, Appendix). Notably, GBM-Stack still led to the lowest biased ATE estimates across the models in our simulations using the adjusted formula. However, the magnitude of bias and the mean squared error of the ATE estimation was higher throughout all considered models, compared to applying Formula 4. We thus recommend the usage of Formula 4 to estimate the ATE (Lunceford and Davidian, 2004).

We also investigated the effect of weight truncation, as discussed by Lee et al. (2011) and Austin and Stuart (2015), by setting weights higher than 10 or 20 to the respective threshold. Note that the bias of our stacked models increased marginally with lower weight thresholds, whereas the variability of the point estimates decreased slightly. Finding an optimal truncation threshold is not straightforward, as already mentioned by Lee et al. (2011). Thus, we did not further consider weight truncation in our analysis. Ju et al. (2019) propose the use of adaptive propensity score truncation and Li et al. (2018) suggest the use of overlap weights, which might further improve the performance of our models.

As an alternative to IPTW, covariate adjustment using the propensity score has been applied (Austin, 2010) to estimate the ATE, by adding the estimated propensity scores as a variable in the outcome regression model, regressing the outcome variable on the binary treatment status. Preliminary results in our simulation study suggest that GBM-Stack, LR-Stack, and Superlearner outperform the other examined propensity score estimators, leading to bias reduction comparable to IPTW and lower variability in the ATE estimation. However, a major drawback of this approach is that the design and analysis of the study are not separated (Austin, 2011). One has to determine the correct relation between the treatment and the outcome (e.g., linear or non-linear), which we assumed to know in our examinations. Further comparisons, considering misspecification of the outcome model might be necessary for a full assessment of the method. For the interested reader, we refer to Autenrieth (2018) for more details of our preliminary investigations discussed in this section.

In addition to estimating the ATE, improving estimation of average treatment effect in the treated group (ATT) on observational, educational data by means of our stacked ensembles, is of great interest to us. We consider a comprehensive analysis of stacked ensemble propensity score matching methods (Caliendo and Kopeinig, 2008; Dehejia and Wahba, 2002) to estimate the ATT, and IPTW using ATT weights, in ongoing work. Eventually, we suppose that our proposed stacked generalization propensity score estimation method could also be successfully applied to obtain unbiased treatment effects in observational studies beyond educational data, such as (precision) medicine, psychology, and econometrics among others.

References

Alcott, B. 2017. Does teacher encouragement influence students' educational progress? A propensity-score matching analysis. *Research in Higher Education 58,* 7, 773–804.

Ali, M. S., Groenwold, R. H., Pestman, W. R., Belitser, S. V., Roes, K. C., Hoes, A. W., de Boer, A., and Klungel, O. H. 2014. Propensity score balance measures in pharmacoepidemiology: A simulation study. *Pharmacoepidemiology and Drug Safety 23,* 8, 802–811.

Alpaydin, E. 2014. *Introduction to Machine Learning*. MIT Press, Cambridge, MA.

Austin, P. C. 2010. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine 29,* 20, 2137–2148.

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research 46,* 3, 399–424.

Austin, P. C. and Stuart, E. A. 2015. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine 34,* 28, 3661–3679.

Autenrieth, M. 2018. Ensemble learning for propensity score methods on in observational studies. [Master's thesis, San Diego State University], ProQuest Dissertations and Theses.

Bakker, T. C., Krabbendam, L., Bhulai, S., and Begeer, S. 2020. First-year progression and retention of autistic students in higher education: A propensity score-weighted population study. *Autism in Adulthood 2,* 4, 307–316.

BEEMER, J., SPOON, K., HE, L., FAN, J., AND LEVINE, R. A. 2018. Ensemble learning for estimating individualized treatment effects in student success studies. *International Journal of Artificial Intelligence in Education 28,* 3, 315–335.

BRAND, J. E. AND XIE, Y. 2010. Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review 75,* 2, 273–302.

BREIMAN, L. 1996. Stacked regressions. *Machine Learning 24,* 1, 49–64.

BUJA, A., STUETZLE, W., AND SHEN, Y. 2005. Loss functions for binary class probability estimation and classification: Structure and applications. Tech. rep., University of Pennsylvania.

CALIENDO, M. AND KOPEINIG, S. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys 22,* 1, 31–72.

CARUANA, E., CHEVRET, S., RESCHE-RIGON, M., AND PIRRACCHIO, R. 2015. A new weighted balance measure helped to select the variables to be included in a propensity score model. *Journal of Clinical Epidemiology 68,* 12, 1415–1422.

CLARK, M. AND CUNDIFF, N. L. 2011. Assessing the effectiveness of a college freshman seminar using propensity score adjustments. *Research in Higher Education 52,* 6, 616–639.

COLE, S. R. AND HERNÁN, M. A. 2008. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology 168,* 6, 656–664.

DEHEJIA, R. H. AND WAHBA, S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics 84,* 1, 151–161.

DOUMPOS, M. AND ZOPOUNIDIS, C. 2007. Model combination for credit risk assessment: A stacked generalization approach. *Annals of Operations Research 151,* 1, 289–306.

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

FEILD, J. L., LEWKOW, N., ZIMMERMAN, N. L., RIEDESEL, M., AND ESSA, A. 2016. A scalable learning analytics platform for automated writing feedback. In *Proceedings of the 9th International Conference on Educational Data Mining*, T. Barnes, M. Chi, and M. Feng, Eds. International Educational Data Mining Society, 688–693.

FRANKLIN, J. M., RASSEN, J. A., ACKERMANN, D., BARTELS, D. B., AND SCHNEEWEISS, S. 2014. Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine 33,* 10, 1685–1699.

FRIEDMAN, J. H. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29,* 5, 1189–1232.

GOLINELLI, D., RIDGEWAY, G., RHOADES, H., TUCKER, J., AND WENZEL, S. 2012. Bias and variance trade-offs when combining propensity score weighting and regression: With an application to HIV status and homeless men. *Health Services and Outcomes Research Methodology 12,* 2-3, 104–118.

GRIFFIN, B. A., MCCAFFREY, D. F., ALMIRALL, D., BURGETTE, L. F., AND SETODJI, C. M. 2017. Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of Causal Inference 5,* 2, 1–18.

GUARCELLO, M. A., LEVINE, R. A., BEEMER, J., FRAZEE, J. P., LAUMAKIS, M. A., AND SCHELLENBERG, S. A. 2017. Balancing student success: Assessing supplemental instruction through coarsened exact matching. *Technology, Knowledge and Learning 22,* 3, 335–352.

HAINMUELLER, J. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20,* 1, 25–46.

Harder, V. S., Stuart, E. A., and Anthony, J. C. 2010. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods 15,* 3, 234–249.

Hirano, K., Imbens, G. W., and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71,* 4, 1161–1189.

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

Imai, K. and Van Dyk, D. A. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association 99,* 467, 854–866.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86,* 1, 4–29.

Jiang, F. and McComas, W. F. 2015. The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education 37,* 3, 554–576.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. 2004. Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician 58,* 4, 272–279.

Ju, C., Schwab, J., and van der Laan, M. J. 2019. On adaptive propensity score truncation in causal inference. *Statistical Methods in Medical Research 28,* 6, 1741–1760.

Kam, C. D. and Palmer, C. L. 2008. Reconsidering the effects of education on political participation. *The Journal of Politics 70,* 3, 612–631.

Kauermann, G. and Carroll, R. J. 2001. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association 96,* 456, 1387–1396.

Kim, R. H. and Clark, D. 2013. The effect of prison-based college education programs on recidivism: Propensity score matching approach. *Journal of Criminal Justice 41,* 3, 196–204.

Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software, Articles 28,* 5, 1–26.

Leblanc, M. and Tibshirani, R. 1996. Combining estimates in regression and classification. *Journal of the American Statistical Association 91,* 436, 1641–1650.

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

Lee, B. K., Lessler, J., and Stuart, E. A. 2011. Weight trimming and propensity score weighting. *Plos One 6,* 3, e18174.

Li, F., Morgan, K. L., and Zaslavsky, A. M. 2018. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association 113,* 521, 390–400.

Lumley, T. 2004. Analysis of complex survey samples. *Journal of Statistical Software 9,* 1, 1–19.

Lunceford, J. K. and Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine 23,* 19, 2937–2960.

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, 403–425.

Microsoft Corporation and Weston, S. 2017. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.11.

Morgan, P. L., Frisco, M. L., Farkas, G., and Hibel, J. 2010. A propensity score matching analysis of the effects of special education services. *The Journal of Special Education 43,* 4, 236–254.

Naimi, A. I. and Balzer, L. B. 2018. Stacked generalization: An introduction to super learning. *European Journal of Epidemiology 33,* 5, 459–464.

Parast, L., McCaffrey, D. F., Burgette, L. F., de la Guardia, F. H., Golinelli, D., Miles, J. N. V., and Griffin, B. A. 2017. Optimizing variance-bias trade-off in the TWANG package for estimation of propensity scores. *Health Services and Outcomes Research Methodology 17,* 3, 175–197.

Pardos, Z. A., Gowda, S. M., Baker, R. S., and Heffernan, N. T. 2012. The sum is greater than the parts: Ensembling models of student knowledge in educational software. *ACM SIGKDD Explorations Newsletter 13,* 2, 37–44.

Pelaez, K., Levine, R., Fan, J., Guarcello, M., and Laumakis, M. 2019. Using a latent class forest to identify at-risk students in higher education. *Journal of Educational Data Mining 11,* 1, 18–46.

Pirracchio, R. and Carone, M. 2018. 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 27,* 8, 2504–2518.

Pirracchio, R., Carone, M., Rigon, M. R., Caruana, E., Mebazaa, A., and Chevret, S. 2016. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Statistical Methods in Medical Research 25,* 5, 1938–1954.

Pirracchio, R., Petersen, M. L., and van der Laan, M. 2014. Improving propensity score estimators' robustness to model misspecification using Super Learner. *American Journal of Epidemiology 181,* 2, 108–119.

Polikar, R. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine 6,* 3, 21–45.

Polikar, R. 2007. Bootstrap - inspired techniques in computation intelligence. *IEEE Signal Processing Magazine 24,* 4, 59–72.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridgeway, G., McCaffrey, D. F., Morral, A. R., Burgette, L. F., and Griffin, B. A. 2014. Toolkit for weighting and analysis of nonequivalent groups. Tech. rep., RAND Corporation.

Rojewski, J. W., Lee, I. H., and Gregg, N. 2015. Causal effects of inclusion on postsecondary education outcomes of individuals with high-incidence disabilities. *Journal of Disability Policy Studies 25,* 4, 210–219.

Rosenbaum, P. R. 1987. Model-based direct adjustment. *Journal of the American Statistical Association 82,* 398, 387–394.

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. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79,* 387, 516–524.

Rubin, D. B. 2001. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology 2,* 3, 169–188.

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

Sesmero, M. P., Ledezma, A. I., and Sanchis, A. 2015. Generating ensembles of heterogeneous classifiers using stacked generalization. *WIREs Data Mining and Knowledge Discovery 5,* 1, 21–34.

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.

Shapiro, J. and Trevino, J. M. 2004. Compensatory education for disadvantaged mexican students: An impact evaluation using propensity score matching. Policy Research Working Paper WPS3334, World Bank.

Stone, C. A. and Tang, Y. 2013. Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research, and Evaluation 18,* 13, 1–12.

Stuart, E. A., Lee, B. K., and Leacy, F. P. 2013. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology 66,* 8, S84–S90.

Sullivan, A. L. and Field, S. 2013. Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis. *Journal of School Psychology 51,* 2, 243–260.

Ting, K. M. and Witten, I. H. 1999. Issues in stacked generalization. *Journal of Artificial Intelligence Research 10,* 1 (May), 271–289.

Titus, M. A. 2007. Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. *Research in Higher Education 48,* 4, 487–521.

Torres-Sospedra, J., Hernández-Espinosa, C., and Fernández-Redondo, M. 2006. Combining MF networks: A comparison among statistical methods and stacked generalization. In *Artificial Neural Networks in Pattern Recognition*, F. Schwenker and S. Marinai, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 210–220.

UMKC 2018. Supplemental intruction. https://info.umkc.edu/si/, accessed July 2018.

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

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.

Williamson, E., Morley, R., Lucas, A., and Carpenter, J. 2012. Propensity scores: From naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research 21,* 3, 273–293.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks 5,* 2, 241 – 259.

Zadrozny, B. and Elkan, C. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, C. E. Brodley and A. P. Danyluk, Eds. Morgan Kaufmann, 609–616.

8. Appendix A - Simulation Study

Scenario A (a model with additivity and linearity)

$$P[Z = 1|X_i] = (1 + \exp\{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7)\})^{-1} \tag{9}$$

Scenario B (a model with mild non-linearity)

$$\begin{aligned} P[Z = 1|X_i] = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_8 X_2^2)\})^{-1} \end{aligned}$$

Scenario C (a model with moderate non-linearity)

$$\begin{aligned} P[Z = 1|X_i] = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_8 X_2^2 + \beta_9 X_4^2 + \beta_{10} X_7^2)\})^{-1} \end{aligned}$$

Scenario D (a model with mild non-additivity)

$$\begin{aligned} P[Z = 1|X_i] = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_8 X_1 X_3 + \beta_9 X_2 X_4 + \beta_{10} X_4 X_5 + \beta_{11} X_5 X_6)\})^{-1} \end{aligned}$$

Scenario E (a model with mild non-additivity and non-linearity)

$$\begin{aligned} P[Z = 1|X_i] = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_8 X_2^2 \\ &+ \beta_9 X_1 X_3 + \beta_{10} X_2 X_4 + \beta_{11} X_4 X_5 + \beta_{12} X_5 X_6)\})^{-1} \end{aligned}$$

Scenario F (a model with moderate non-additivity)

$$\begin{aligned} P[Z = 1|X_i] = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_8 X_1 X_3 + \beta_9 X_2 X_4 + \beta_{10} X_3 X_5 + \beta_{11} X_4 X_6 + \beta_{12} X_5 X_7 \\ &+ \beta_{13} X_1 X_6 + \beta_{14} X_2 X_3 + \beta_{15} X_4 X_5 + \beta_{16} X_5 X_6)\})^{-1} \end{aligned}$$

Scenario G (a model with moderate non-additivity and non-linearity)

$$\begin{aligned} P[Z = 1|X_i] = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_8 X_2^2 + \beta_9 X_4^2 + \beta_{10} X_7^2 \\ &+ \beta_{11} X_1 X_3 + \beta_{12} X_2 X_4 + \beta_{13} X_3 X_5 + \beta_{14} X_4 X_6 + \beta_{15} X_5 X_7 \\ &+ \beta_{16} X_1 X_6 + \beta_{17} X_2 X_3 + \beta_{18} X_3 X_4 + \beta_{19} X_4 X_5 + \beta_{20} X_5 X_6\})^{-1} \end{aligned}$$

Scenario H (a model with severe non-additivity and non-linearity)

$$\begin{aligned} P[Z = 1|X_i] = (1 + \exp\{-(&\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_8 X_2^2 + \beta_9 \sqrt{|X_4|} + \beta_{10} X_7^3 \\ &+ \beta_{11} X_1 X_3 + \beta_{12} X_2 X_4 + \beta_{13} X_3 X_7 \\ &+ \beta_{14} X_4 X_6 + \beta_{15} X_4 X_7 + \beta_{16} X_2 X_4 \\ &+ \beta_{17} X_2 X_3 X_7 + \beta_{18} X_2 X_4 X_7 + \beta_{19} X_4 X_5 X_2 + \beta_{20} X_5 X_6 X_7\})^{-1} \end{aligned}$$

Outcome model: Scenario A-H

$$\begin{aligned} P[Y = 1|Z, X_i] = (1 + \exp\{-(&\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 \\ &+ \alpha_6 X_9 + \alpha_7 X_{10} + \gamma_{LO} Z\})^{-1} \end{aligned}$$

Table 4: Description of base learner methods used in our Stacked Ensemble.

Model Description			
Model	**Method**	**Tuning Parameters**	**Libraries**
Random Forest (RF): Consists of several decision trees. For each independent variable, the data set is split at several split points. The sum of squared error is calculated at each split point. The variable resulting in minimum SSE is selected for the node. This process recursively continues until the tree reaches a stopping criteria, usually a terminal node size. The process is done for a large number of decision trees.	rf	mtry	randomForests
Stochastic Gradient Boosting (GBM): Constructs additive regression models by sequentially fitting a simple parameterized function to current pseudo-residuals by least squares at each iteration. The pseudo-residuals are the gradient of the loss function minimized. At each iteration, a subsample of the training data is drawn at random without replacement from the training data. This is used in place of the full sample to fit the base learner and compute the model update for the current iteration. Increases robustness against overcapacity of the base learner.	gbm	n.trees, interaction.depth, shrinkage, n.minobsinnode	gbm, plyr

Model Description			
Model	**Method**	**Tuning Parameters**	**Libraries**
Neural Network (NNET): Used to estimate functions that can depend on a large number of inputs and are generally unknown. They are typically organized layers. Layers are made up of a number of interconnected nodes which contain an activation function. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the processing is done via a system of weighted connections. The hidden layers kink then to an output layer where the answer is the output.	nnet	size, decay	nnet
Bagged CART(BAG-CART): This is a 'perturb and combine' method. We draw B bootstrap samples. A bootstrap sample is a random sample of size n drawn from the empirical distribution of a sample of size n; that is, the training data is resampled with replacement. Some of the cases will be left out of the sample and some cases will be represented more than once. Then, build a tree on each bootstrap sample. Pruning can be counterproductive. Large trees with low bias and high variance are ideal. The final aggregate classifier can be obtained by majority voting for classification.	treebag	None	ipred, plyr, e1071
Support Vector Machines with Radial Basis Function Kernel (SVM): Builds a model that assigns new examples to one category or the other. Is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall in. SVM can perform a linear classification as well as a nonlinear classification.	svmRadial	sigma, C	kernlab

Model Description			
Model	**Method**	**Tuning Parameters**	**Libraries**
Regularized Logistic Regression (RLR):A logistic regression is a special form of linear regression where the dependent variable is categorical. The regularized logistic regression is a logistic regression with a regularized term	regLogistic	cost, loss, epsilon	LiblineaR
Boosted Logistic Regression (BOOSTLR): This method is a boosting method with a logistic regression as cost function.	LogitBoost	nIter	caTools
Averaged Neural Network(AVNNET): A neural network ensemble. The neural network model is fit using different random number seeds. All the resulting models are used for prediction. For classification, the model scores are first averaged, then translated to predicted classes. Bagging can also be used to create the models.	avNNet	size, decay, bag	nnet

Model Description			
Model	**Method**	**Tuning Parameters**	**Libraries**
Nearest Shrunken Centroids (SC): Nearest centroid classification takes the gene expression profile of a new sample and compares it to each of these class centroids. The class whose centroid that it is closest to, in squared distance, is the predicted class for that sample. The shrunken centroid "shrinks" each of the class centroids toward the overall centroid for all classes by an amount we call the threshold by moving the centroid towards zero by threshold, setting it equal to zero if it hits zero.	pam	threshold	pamr
Generalized Linear Model (LR): Logistic Regression	glm	None	
k-Nearest Neighbors (KNN): A non-parametric method. The input consists of the k closest training examples in the feature space. The output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.	kknn	kmax, distance, kernel	kknn
Naive Bayes (NB): Family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. These classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.	nb	fL, usekernel, adjust	klaR

Table 5: Linear combinations and introduced correlations between simulation covariates.

Variables	Confounders				Instrumental Variable			Outcome Predictors		
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$X_1 = V_1 + \sqrt{\frac{0.2^2}{(1-0.2^2)}} * V_5$	1									
$X_2 = V_2 + \sqrt{\frac{0.9^2}{(1-0.9^2)}} * V_6$	0	1								
$X_3 = V_3 + \sqrt{\frac{0.2^2}{(1-0.2^2)}} * V_8$	0	0	1							
$X_4 = V_4 + \sqrt{\frac{0.9^2}{(1-0.9^2)}} * V_9$	0	0	0	1						
$X_5 = V_5$	0.2	0	0	0	1					
$X_6 = V_6$	0	0.9	0	0	0	1				
X_7	0	0	0	0	0	0	1			
$X_8 = V_8$	0	0	0.2	0	0	0	0	1		
$X_9 = V_9$	0	0	0	0.9	0	0	0	0	1	
X_{10}	0	0	0	0	0	0	0	0	0	1

Table 6: Parameters for Outcome Model.

Model	α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7	γ_{LO}
A-H	-1.9	0.3	-0.36	-0.73	-0.2	0.71	-0.19	0.26	-0.8

Table 7: Parameters for Treatment Effect Model.

Model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}
A	-1.2	0.3	-0.5	0.6	-0.3	0.7	-0.6	0.5													
B	-1.3	0.3	-0.5	0.6	-0.3	0.7	-0.6	0.5	0.5												
C	-1.2	-0.6	0.3	0.6	-0.3	0.7	-0.6	0.5	0.4	0.5	0.7										
D	-1.2	0.3	-0.5	0.6	-0.3	0.7	-0.6	0.5	1	0.6	-0.6	-0.8									
E	-1.3	0.3	-0.5	0.6	-0.3	0.7	-0.6	0.5	0.4	1	0.4	-0.4	-0.8								
F	-1.2	0.3	-0.5	0.6	-0.3	0.7	-0.6	0.5	1	0.6	0.6	-0.6	-0.8	1	0.4	-0.6	-0.8				
G	-1.3	-0.6	0.3	0.6	-0.3	0.7	-0.6	0.5	0.4	-0.5	0.7	0.5	0.2	0.6	-0.4	-0.8	1	0.4	0.6	-0.4	-0.56
H	-1.3	0.3	-0.5	0.6	-0.3	0.7	-0.6	0.5	0.4	-0.5	-0.3	0.5	0.2	0.6	-0.4	-0.4	0.5	0.4	0.6	-0.4	-0.3

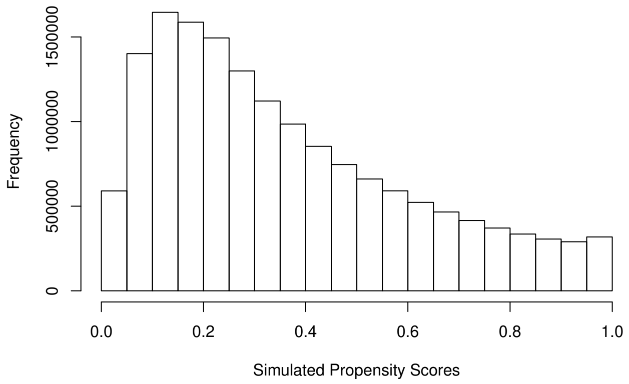

Figure 4: Histogram of the collective simulated propensity scores over all scenarios and all data sets.

9. Appendix A - Additional formulas

Log-loss: The log-loss or cross-entropy loss defined by

$$-\frac{1}{n}\sum_{i=1}^{n}\{y_i \log p_i + (1-y_i)\log(1-p_i)\}\,, \tag{10}$$

has shown high accordance with prediction errors in class probability estimates (Zadrozny and Elkan, 2001; Buja et al., 2005). In Formula (10), n is the sample size, y_i is the true class label and p_i the predicted class probability, the propensity score, for observation $i \in \{1, \ldots n\}$.

Alternative average treatment effect estimator: according to Austin (2011)

$$ATE_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_iY_i}{e_i(X)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e_i(X)}, \tag{11}$$

Average treatment effect in the actual treated group (ATT): The weights to estimate the ATT are defined by

$$w_{ATT} = Z + \frac{e(X)(1-Z)}{1-e(X)}. \tag{12}$$

10. Appendix A – ASAM balance literature

In this section, we introduce a brief review about recommendations and usage of standardized mean differences, particularly the ASAM, as a balance measure in the literature, in which we contextualize our results in Section 4

Franklin et al. (2014), Stuart et al. (2013) and Ali et al. (2014) indicate a positive relationship between the decrease of ASAM and the reduction of bias in the treatment effect estimate in their simulation studies. Ridgeway et al. (2014) and Pirracchio and Carone (2018) use ASAM as a loss function for their propensity score estimation model fit, selecting model hyperparameters to optimize ASAM. However, in the literature, there is no accordance regarding a critical magnitude of absolute standardized mean difference between covariates. Austin and Stuart (2015) suggest a threshold of 0.1, Lee et al. (2010) of 0.2 and Harder et al. (2010) consider values greater than 0.25 as concerning. Besides, Lee et al. (2010) and Pirracchio et al. (2014) point out that lower ASAM does not always lead to lower bias, specifically considering logistic regression-based estimates. Golinelli et al. (2012) found that optimization of balance, in general, could increase the variance of the point estimate at some level. Further concerns have been raised regarding the use of the ASAM in the presence of instrumental variables, e.g., (Pirracchio and Carone, 2018; Caruana et al., 2015). By just comparing the ASAM, one does not distinguish between the decrease of imbalance in predictive important confounders and not outcome associated instrumental variables. This could lead to further variation in the point estimate. Nevertheless, in recent simulation studies assessing the performance of different propensity score estimation methods, the averaged standardized absolute mean differences (ASAM) are evaluated to check the balancing ability of the introduced methods (Lee et al., 2010; Pirracchio et al., 2014; Pirracchio et al., 2016).

11. Appendix A – Additional ASAM balance results

11.1. Appendix A – ASAM of confounding covariates

In addition to ASAM (which represents the averaged standardized absolute mean differences of all covariates), we also computed the average of the standardized absolute mean differences including the four confounding covariates only, hereafter called ASAM$_{conf}$. For applied researchers, the ASAM$_{conf}$ is usually not directly accessible since normally one cannot distinguish between confounders, instrumental, or just outcome-related covariates. Nevertheless, the balance assessment of the ASAM$_{conf}$ values is of particular interest, since the confounding covariates are both highly unbalanced due to their association to the treatment and highly influential in the treatment effect estimate due to their strong relation to the outcome. Superlearner and GBM-Stack led to the lowest ASAM$_{conf}$ values with an average of 0.101 and 0.102 across the setups, again similar to the ASAM$_{conf}$ of the simulated "true" propensity score (0.103). Note that these magnitudes are only marginally higher than the respective ASAM (including all covariates), thus indicating well-balanced confounders. Twang-GBM had a bit higher ASAM$_{conf}$ (0.108) compared to the ASAM (0.100). Twang-GBM is trained to balance covariates. However, by including non-confounders in the model, which is realistic in an educational set-up, the ASAM is minimized in the possible cost of more imbalance in the set of the important confounders (ASAM$_{conf}$), which could explain the observed performance. The ASAM and ASAM$_{conf}$ for all individual scenarios are listed in the Appendix (Table A.36, 70, 104 and Table A.37, 71, 105).

11.2. Appendix A – Critical magnitude of standardized absolute mean differences

We did not find any indication of a critical magnitude of standardized absolute mean differences in our simulations. Firstly, our simulations suggest that the ASAM is generally lower for larger data setups, which was also stated by Lee et al. (2010). Thus, we suppose that, if existent, a critical value of standardized absolute mean differences cannot be fixed, but should depend on the data size. For either of our data size setups, a critical value of 0.1 was obviously too small to detect meaningful confounding. For instance, our stacked models showed low biased estimates with an averaged ASAM of around 0.1. Besides, on average over the $m = 1000$ simulations, there were between 1.3 (GBM-Stack) and 2.0 (NNET) covariates with standardized absolute mean differences greater than 0.2 in each scenario (Appendix, Table A.122). Further, no Spearman correlations were given between the number of covariates with standardized absolute mean differences greater than 0.2 and the absolute ATE estimation error. Thus, our simulations did not affirm any threshold suggested in the literature (as discussed in Section 10, Appendix) that would generally imply concerning confounding and bias in the point estimate.

12. Appendix A – Simulation study figures

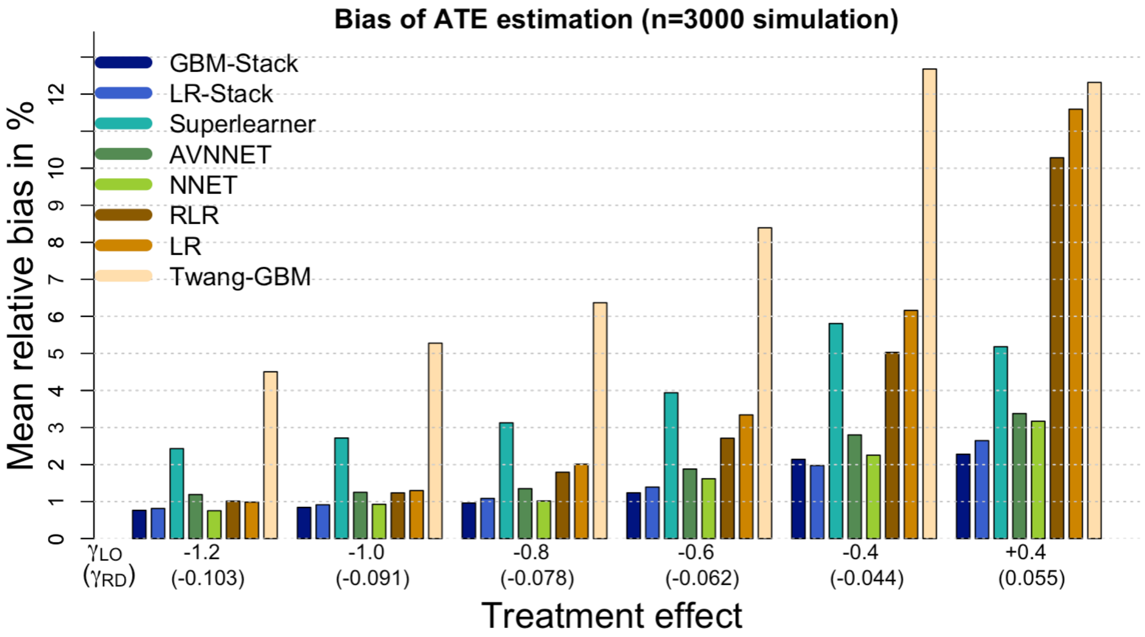

Figure 5: Comparison of relative bias (in %) resulting from the different ATE estimator for data size n=3000. The illustrated bias results present the average bias over the eight scenarios (A-H) for each treatment effect $\gamma_{LO} \in \{-1.2, -1.0, -0.8, -0.6, -0.4, +0.4\}$, respectively.

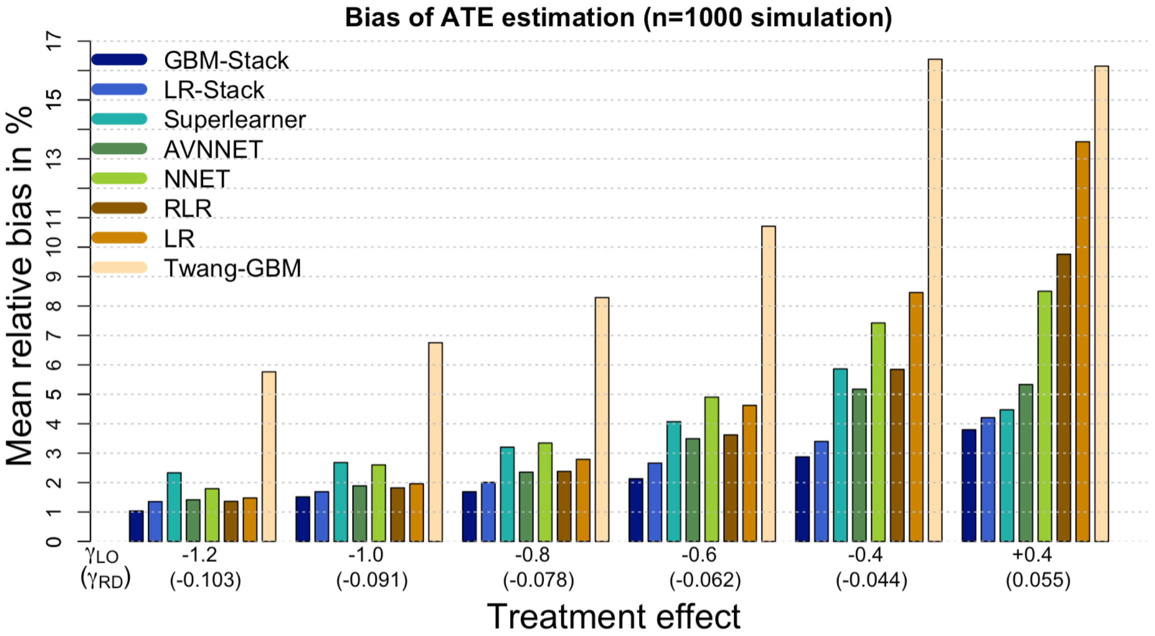

Figure 6: Comparison of relative bias (in %) resulting from the different ATE estimator for data size n=1000. The illustrated bias results present the average bias over the eight scenarios (A-H) for each treatment effect $\gamma_{LO} \in \{-1.2, -1.0, -0.8, -0.6, -0.4, +0.4\}$, respectively.

12.1. Appendix A – Performance measure results based on n = 3000 setup

Note that there was no visible pattern in the direction of the bias. For all learners, there were setups in which they underestimate and setups in which they overestimate the average treatment effects. For a better comparison, we thus only present the absolute values of the relative bias.

Table 8: (Relative Bias; n=3000; ATE= -0.103) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=3000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	1.80	1.38	0.16	0.66	2.83	0.63	0.33	1.73	1.19
BAG-CART	11.94	17.30	2.02	15.72	22.46	12.83	1.32	12.96	12.07
BOOSTLR	9.47	1.17	0.19	4.83	1.54	9.41	7.65	4.43	4.84
GBM	0.90	1.58	2.45	2.25	3.32	1.99	0.63	1.11	1.78
GBM-Stack	0.10	0.35	1.43	0.11	1.38	1.46	0.45	0.83	0.76
KNN	8.20	10.94	3.26	11.60	9.46	10.32	7.45	3.33	8.07
LR	0.15	1.72	0.88	0.90	0.65	0.59	2.16	0.83	0.99
LR-Stack	0.71	1.27	0.47	0.11	1.79	1.12	0.04	1.00	0.81
NB	15.04	9.08	3.37	21.12	13.01	16.26	1.45	5.90	10.65
NNET	0.72	0.26	0.41	0.97	0.59	1.10	1.68	0.31	0.75
RF	1.33	1.07	1.87	1.29	0.36	1.12	0.11	0.61	0.97
RLR	0.44	2.00	0.66	0.45	1.08	1.05	1.99	0.43	1.01
SC	7.97	9.19	4.56	10.42	10.69	9.66	4.61	8.38	8.19
Superlearner	1.40	2.04	2.10	2.79	3.96	3.36	0.65	3.13	2.43
SVM	2.76	0.31	0.37	0.44	0.09	2.58	1.67	1.09	1.16
Twang-GBM	4.02	5.06	2.36	5.88	6.73	5.37	1.44	5.18	4.51
Simulated PS	0.34	0.09	0.57	0.47	0.12	0.53	0.09	1.63	0.48

Table 9: (Relative Bias; n=3000; ATE= -0.091) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=3000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	2.09	1.46	0.28	0.51	3.04	0.45	0.49	1.68	1.25
BAG-CART	15.63	20.11	3.42	19.68	26.48	16.07	2.64	16.77	15.10
BOOSTLR	10.39	1.68	0.22	6.28	1.46	10.97	8.47	5.82	5.66
GBM	0.98	1.74	3.10	2.53	3.61	2.47	0.69	0.98	2.01
GBM-Stack	0.02	0.27	1.83	0.15	1.36	1.77	0.63	0.71	0.84
KNN	10.64	13.35	4.92	14.84	12.18	13.48	9.33	4.50	10.40
LR	0.07	1.43	1.39	2.09	0.35	0.47	2.75	1.83	1.30
LR-Stack	0.88	1.30	0.68	0.22	1.82	1.42	0.08	0.91	0.91
NB	18.15	11.10	4.14	27.13	16.59	19.76	1.85	8.67	13.42
NNET	0.73	0.09	0.28	1.40	0.53	1.55	1.90	0.93	0.93
RF	2.17	1.67	2.10	2.34	0.38	1.44	0.33	1.19	1.45
RLR	0.42	1.78	1.11	1.49	0.19	1.05	2.52	1.32	1.24
SC	9.47	10.58	5.42	12.51	12.47	11.56	5.62	9.76	9.68
Superlearner	1.58	2.19	2.64	3.10	4.27	3.97	0.69	3.29	2.72
SVM	3.42	0.20	0.26	0.28	0.17	3.21	2.13	1.07	1.34
Twang-GBM	4.70	5.81	2.98	6.94	7.72	6.43	1.70	5.92	5.28
Simulated PS	0.33	0.07	0.31	0.55	0.02	0.72	0.08	1.51	0.45

Table 10: (Relative Bias; n=3000; ATE= -0.078) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=3000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	2.30	1.75	0.57	0.56	2.92	0.20	0.39	2.10	1.35
BAG-CART	21.15	23.98	3.12	24.59	33.35	22.10	2.88	19.02	18.78
BOOSTLR	12.75	2.42	0.02	8.65	1.10	14.07	10.82	7.13	7.12
GBM	0.97	2.07	3.87	2.80	3.72	2.92	0.90	1.50	2.34
GBM-Stack	0.31	0.30	2.21	0.09	0.99	2.04	0.76	0.99	0.96
KNN	14.15	16.38	6.67	19.83	17.36	17.24	11.21	5.96	13.60
LR	0.24	1.05	1.77	3.92	2.16	0.01	3.73	3.19	2.01
LR-Stack	1.37	1.52	0.86	0.42	1.54	1.67	0.01	1.29	1.09
NB	23.34	13.98	5.23	36.48	22.26	25.52	2.01	11.95	17.59
NNET	0.58	0.06	0.02	1.43	0.01	2.43	2.18	1.43	1.02
RF	3.65	2.28	2.74	3.81	1.82	3.23	0.33	1.23	2.38
RLR	0.22	1.52	1.41	3.11	1.41	0.77	3.41	2.49	1.79
SC	11.67	12.87	6.97	15.38	14.87	14.22	7.17	11.96	11.89
Superlearner	1.67	2.54	3.32	3.51	4.41	4.60	0.93	4.02	3.13
SVM	4.32	0.21	0.23	0.17	0.84	4.06	2.43	1.55	1.73
Twang-GBM	5.61	7.02	3.73	8.33	8.93	7.74	2.13	7.45	6.37
Simulated PS	0.07	0.06	0.27	0.54	0.56	0.96	0.34	2.20	0.62

Table 11: (Relative Bias; n=3000; ATE= -0.062) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=3000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	3.31	1.92	1.10	0.45	3.78	0.95	0.62	2.91	1.88
BAG-CART	29.13	30.63	3.24	32.66	42.91	32.03	6.12	26.73	25.43
BOOSTLR	15.61	3.73	0.74	12.12	0.06	20.49	15.06	9.73	9.69
GBM	1.70	2.36	5.51	3.57	4.68	3.12	1.01	2.50	3.06
GBM-Stack	0.06	0.07	3.25	0.27	1.32	1.84	1.20	1.89	1.24
KNN	20.31	22.77	9.78	26.00	24.63	24.91	15.25	8.88	19.07
LR	0.30	0.22	2.50	6.86	4.72	1.50	5.83	4.79	3.34
LR-Stack	1.51	1.35	1.39	0.94	2.00	1.48	0.19	2.28	1.39
NB	32.15	19.79	7.14	51.61	32.03	35.89	2.85	18.17	24.95
NNET	0.87	0.37	0.13	2.38	0.16	4.43	2.93	1.64	1.62
RF	5.06	3.70	3.62	6.10	3.10	5.91	0.29	2.10	3.74
RLR	0.35	0.45	2.01	5.67	3.65	0.40	5.37	3.80	2.71
SC	16.23	16.42	9.61	20.42	19.47	18.36	9.23	16.50	15.78
Superlearner	2.54	2.75	4.66	4.24	5.51	5.31	0.95	5.56	3.94
SVM	6.85	0.25	0.17	0.07	1.00	4.88	3.40	2.79	2.43
Twang-GBM	8.10	8.97	5.27	10.76	11.72	9.40	2.53	10.38	8.39
Simulated PS	0.24	0.20	0.16	0.97	0.35	1.86	0.69	3.31	0.97

Table 12: (Relative Bias; n=3000; ATE= -0.044) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=3000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	5.22	2.42	1.07	1.34	5.16	1.46	0.82	4.92	2.80
BAG-CART	42.42	46.23	1.45	50.29	63.78	48.29	8.71	38.14	37.41
BOOSTLR	22.04	5.26	0.24	18.29	1.91	33.15	23.37	14.23	14.81
GBM	2.78	3.27	7.89	6.30	6.18	5.66	1.66	4.63	4.79
GBM-Stack	0.33	0.65	4.10	1.67	1.24	3.32	2.08	3.73	2.14
KNN	29.19	37.43	14.63	37.77	38.83	40.79	24.23	13.72	29.58
LR	0.23	2.17	4.82	11.40	9.84	2.69	9.78	8.35	6.16
LR-Stack	1.65	1.51	1.28	0.60	2.46	3.09	0.46	4.72	1.97
NB	49.02	30.80	10.15	81.31	51.13	55.41	3.26	31.12	39.02
NNET	1.51	0.88	0.80	2.65	0.05	6.09	3.62	2.43	2.25
RF	7.93	7.34	4.99	8.80	6.66	9.69	0.70	3.50	6.20
RLR	0.77	1.12	4.03	9.45	8.14	0.92	9.04	6.76	5.03
SC	24.79	23.94	14.16	31.09	27.86	28.13	13.81	25.17	23.62
Superlearner	4.24	3.36	6.31	7.16	7.08	8.38	1.19	8.73	5.81
SVM	11.12	0.60	0.68	1.40	2.10	8.50	5.48	4.78	4.33
Twang-GBM	12.61	13.01	7.39	16.87	16.80	14.85	3.68	16.21	12.68
Simulated PS	0.70	0.61	0.31	0.27	0.47	1.77	1.09	4.82	1.26

Table 13: (Relative Bias; n=3000; ATE= 0.055) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=3000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	6.50	4.61	2.22	1.67	6.35	0.32	3.48	1.88	3.38
BAG-CART	44.22	40.69	2.70	51.81	60.29	50.44	7.25	42.34	37.47
BOOSTLR	18.34	7.49	2.48	25.44	2.74	34.59	18.70	16.28	15.76
GBM	4.67	4.69	5.14	6.12	6.94	6.81	0.18	1.94	4.56
GBM-Stack	1.57	0.51	0.55	2.31	2.10	4.63	4.34	2.21	2.28
KNN	30.86	35.78	22.09	44.12	44.76	45.99	33.05	21.71	34.79
LR	0.93	7.27	10.00	20.26	19.32	2.88	14.66	17.45	11.59
LR-Stack	0.17	3.18	2.70	0.29	3.24	4.79	3.09	3.71	2.65
NB	53.46	34.52	6.71	102.41	61.53	58.52	2.51	48.72	46.05
NNET	2.81	1.28	4.36	2.36	1.66	3.59	4.73	4.56	3.17
RF	8.06	8.13	0.81	13.26	9.85	11.24	4.56	8.97	8.11
RLR	2.05	5.95	9.01	17.81	17.19	1.04	13.74	15.47	10.28
SC	27.27	24.21	11.85	31.69	26.59	27.86	12.33	23.22	23.13
Superlearner	5.63	3.83	3.17	6.17	6.46	9.28	1.21	5.70	5.18
SVM	15.34	0.45	4.41	3.96	1.31	10.86	8.41	4.34	6.13
Twang-GBM	14.72	14.14	4.41	16.82	17.26	15.50	1.36	14.33	12.32
Simulated PS	1.95	1.75	4.19	2.29	1.82	0.74	1.33	4.43	2.31

Table 14: (Mse; n=3000; ATE= -0.103) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0006	0.0007	0.0007	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007
BAG-CART	0.0033	0.0078	0.0064	0.0064	0.0083	0.0066	0.0072	0.0070	0.0066
BOOSTLR	0.0019	0.0016	0.0022	0.0029	0.0023	0.0038	0.0020	0.0020	0.0023
GBM	0.0006	0.0006	0.0007	0.0006	0.0007	0.0007	0.0007	0.0007	0.0007
GBM-Stack	0.0006	0.0006	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
KNN	0.0051	0.0070	0.0045	0.0080	0.0088	0.0081	0.0052	0.0042	0.0064
LR	0.0006	0.0006	0.0005	0.0007	0.0006	0.0007	0.0005	0.0006	0.0006
LR-Stack	0.0006	0.0006	0.0007	0.0007	0.0007	0.0008	0.0008	0.0007	0.0007
NB	0.0011	0.0010	0.0007	0.0016	0.0012	0.0013	0.0007	0.0008	0.0010
NNET	0.0006	0.0008	0.0009	0.0010	0.0009	0.0010	0.0010	0.0010	0.0009
RF	0.0008	0.0007	0.0008	0.0008	0.0009	0.0008	0.0007	0.0009	0.0008
RLR	0.0006	0.0006	0.0005	0.0007	0.0006	0.0007	0.0005	0.0006	0.0006
SC	0.0006	0.0006	0.0005	0.0006	0.0006	0.0006	0.0005	0.0006	0.0006
Superlearner	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0006
SVM	0.0006	0.0006	0.0007	0.0008	0.0008	0.0009	0.0007	0.0007	0.0007
Twang-GBM	0.0006	0.0006	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0006
Simulated PS	0.0006	0.0008	0.0009	0.0010	0.0012	0.0010	0.0012	0.0010	0.0010

Table 15: (Mse; n=3000; ATE= -0.091) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0007	0.0007	0.0008	0.0009	0.0008	0.0008	0.0007	0.0008	0.0008
BAG-CART	0.0036	0.0083	0.0071	0.0067	0.0086	0.0072	0.0075	0.0075	0.0071
BOOSTLR	0.0022	0.0017	0.0024	0.0031	0.0025	0.0041	0.0021	0.0022	0.0025
GBM	0.0007	0.0007	0.0008	0.0007	0.0008	0.0007	0.0008	0.0008	0.0007
GBM-Stack	0.0006	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0007
KNN	0.0055	0.0074	0.0050	0.0084	0.0091	0.0083	0.0057	0.0046	0.0068
LR	0.0007	0.0006	0.0006	0.0008	0.0007	0.0008	0.0006	0.0007	0.0007
LR-Stack	0.0007	0.0006	0.0008	0.0008	0.0008	0.0009	0.0008	0.0008	0.0008
NB	0.0013	0.0010	0.0007	0.0019	0.0013	0.0014	0.0008	0.0010	0.0012
NNET	0.0007	0.0008	0.0009	0.0011	0.0010	0.0011	0.0011	0.0011	0.0010
RF	0.0009	0.0008	0.0009	0.0009	0.0010	0.0010	0.0008	0.0010	0.0009
RLR	0.0007	0.0006	0.0006	0.0008	0.0007	0.0007	0.0006	0.0006	0.0007
SC	0.0006	0.0006	0.0006	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006
Superlearner	0.0006	0.0006	0.0006	0.0007	0.0007	0.0007	0.0006	0.0007	0.0007
SVM	0.0007	0.0007	0.0008	0.0009	0.0008	0.0009	0.0008	0.0008	0.0008
Twang-GBM	0.0006	0.0006	0.0006	0.0007	0.0007	0.0007	0.0006	0.0007	0.0007
Simulated PS	0.0007	0.0009	0.0010	0.0010	0.0013	0.0010	0.0012	0.0011	0.0010

Table 16: (Mse; n=3000; ATE= -0.078) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0008	0.0009	0.0009
BAG-CART	0.0042	0.0089	0.0074	0.0073	0.0093	0.0078	0.0083	0.0081	0.0076
BOOSTLR	0.0024	0.0019	0.0027	0.0034	0.0027	0.0044	0.0024	0.0023	0.0028
GBM	0.0008	0.0007	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008
GBM-Stack	0.0007	0.0007	0.0008	0.0008	0.0008	0.0009	0.0008	0.0008	0.0008
KNN	0.0059	0.0080	0.0055	0.0087	0.0095	0.0087	0.0062	0.0051	0.0072
LR	0.0008	0.0007	0.0006	0.0009	0.0008	0.0008	0.0006	0.0007	0.0008
LR-Stack	0.0008	0.0007	0.0008	0.0008	0.0008	0.0009	0.0008	0.0008	0.0008
NB	0.0015	0.0012	0.0008	0.0022	0.0015	0.0016	0.0009	0.0011	0.0013
NNET	0.0008	0.0009	0.0010	0.0012	0.0011	0.0012	0.0012	0.0012	0.0011
RF	0.0010	0.0009	0.0010	0.0010	0.0011	0.0011	0.0009	0.0012	0.0010
RLR	0.0008	0.0007	0.0006	0.0009	0.0008	0.0008	0.0006	0.0007	0.0007
SC	0.0007	0.0006	0.0006	0.0007	0.0007	0.0007	0.0006	0.0007	0.0007
Superlearner	0.0007	0.0007	0.0007	0.0008	0.0008	0.0008	0.0007	0.0008	0.0007
SVM	0.0008	0.0007	0.0008	0.0010	0.0009	0.0010	0.0008	0.0009	0.0009
Twang-GBM	0.0007	0.0007	0.0007	0.0007	0.0008	0.0007	0.0007	0.0007	0.0007
Simulated PS	0.0008	0.0009	0.0010	0.0011	0.0013	0.0011	0.0013	0.0013	0.0011

Table 17: (Mse; n=3000; ATE= -0.062) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0008	0.0009	0.0008	0.0010	0.0010	0.0010	0.0008	0.0010	0.0009
BAG-CART	0.0048	0.0095	0.0078	0.0080	0.0101	0.0083	0.0091	0.0087	0.0083
BOOSTLR	0.0026	0.0020	0.0029	0.0038	0.0030	0.0047	0.0027	0.0026	0.0030
GBM	0.0008	0.0008	0.0008	0.0008	0.0009	0.0008	0.0008	0.0009	0.0009
GBM-Stack	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0009	0.0009	0.0009
KNN	0.0063	0.0085	0.0059	0.0093	0.0103	0.0093	0.0064	0.0056	0.0077
LR	0.0009	0.0008	0.0007	0.0010	0.0009	0.0009	0.0007	0.0008	0.0008
LR-Stack	0.0008	0.0008	0.0008	0.0009	0.0009	0.0010	0.0009	0.0009	0.0009
NB	0.0017	0.0013	0.0008	0.0027	0.0017	0.0018	0.0009	0.0014	0.0015
NNET	0.0009	0.0010	0.0011	0.0013	0.0012	0.0013	0.0013	0.0014	0.0012
RF	0.0011	0.0011	0.0010	0.0011	0.0013	0.0012	0.0009	0.0014	0.0011
RLR	0.0009	0.0007	0.0007	0.0010	0.0009	0.0009	0.0007	0.0008	0.0008
SC	0.0008	0.0007	0.0006	0.0008	0.0008	0.0007	0.0006	0.0008	0.0007
Superlearner	0.0008	0.0008	0.0007	0.0008	0.0008	0.0008	0.0007	0.0008	0.0008
SVM	0.0009	0.0008	0.0008	0.0010	0.0010	0.0010	0.0009	0.0010	0.0009
Twang-GBM	0.0008	0.0008	0.0007	0.0008	0.0008	0.0008	0.0007	0.0008	0.0008
Simulated PS	0.0008	0.0010	0.0011	0.0012	0.0014	0.0012	0.0014	0.0014	0.0012

Table 18: (Mse; n=3000; ATE= -0.044) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0009	0.0009	0.0009	0.0011	0.0011	0.0011	0.0010	0.0011	0.0010
BAG-CART	0.0054	0.0102	0.0084	0.0086	0.0108	0.0090	0.0098	0.0094	0.0089
BOOSTLR	0.0027	0.0023	0.0032	0.0042	0.0033	0.0050	0.0030	0.0029	0.0033
GBM	0.0009	0.0009	0.0009	0.0009	0.0010	0.0009	0.0009	0.0010	0.0009
GBM-Stack	0.0009	0.0008	0.0009	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
KNN	0.0066	0.0092	0.0065	0.0098	0.0109	0.0099	0.0067	0.0062	0.0082
LR	0.0009	0.0008	0.0007	0.0012	0.0010	0.0010	0.0008	0.0009	0.0009
LR-Stack	0.0009	0.0008	0.0009	0.0010	0.0010	0.0011	0.0010	0.0010	0.0010
NB	0.0019	0.0014	0.0009	0.0032	0.0019	0.0020	0.0010	0.0016	0.0017
NNET	0.0009	0.0010	0.0012	0.0014	0.0013	0.0014	0.0014	0.0014	0.0013
RF	0.0012	0.0012	0.0011	0.0013	0.0014	0.0013	0.0011	0.0015	0.0013
RLR	0.0009	0.0008	0.0007	0.0011	0.0010	0.0010	0.0008	0.0009	0.0009
SC	0.0008	0.0008	0.0006	0.0009	0.0008	0.0008	0.0007	0.0008	0.0008
Superlearner	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0008	0.0009	0.0009
SVM	0.0009	0.0009	0.0009	0.0011	0.0011	0.0011	0.0010	0.0010	0.0010
Twang-GBM	0.0008	0.0008	0.0008	0.0009	0.0009	0.0008	0.0008	0.0009	0.0008
Simulated PS	0.0009	0.0011	0.0012	0.0013	0.0015	0.0013	0.0015	0.0015	0.0013

Table 19: (Mse; n=3000; ATE= 0.055) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0013	0.0013	0.0012	0.0017	0.0016	0.0016	0.0013	0.0016	0.0015
BAG-CART	0.0095	0.0124	0.0113	0.0126	0.0155	0.0131	0.0127	0.0137	0.0126
BOOSTLR	0.0038	0.0031	0.0048	0.0061	0.0049	0.0070	0.0042	0.0046	0.0048
GBM	0.0013	0.0012	0.0012	0.0014	0.0014	0.0013	0.0012	0.0015	0.0013
GBM-Stack	0.0013	0.0011	0.0012	0.0014	0.0014	0.0014	0.0013	0.0014	0.0013
KNN	0.0093	0.0120	0.0090	0.0129	0.0135	0.0128	0.0096	0.0085	0.0110
LR	0.0014	0.0012	0.0010	0.0021	0.0017	0.0015	0.0011	0.0015	0.0014
LR-Stack	0.0013	0.0011	0.0012	0.0014	0.0014	0.0015	0.0014	0.0014	0.0013
NB	0.0031	0.0021	0.0011	0.0068	0.0034	0.0033	0.0014	0.0028	0.0030
NNET	0.0014	0.0014	0.0017	0.0021	0.0018	0.0021	0.0019	0.0021	0.0018
RF	0.0021	0.0018	0.0016	0.0023	0.0022	0.0021	0.0016	0.0023	0.0020
RLR	0.0014	0.0012	0.0010	0.0019	0.0016	0.0014	0.0010	0.0014	0.0014
SC	0.0012	0.0010	0.0008	0.0013	0.0012	0.0011	0.0009	0.0012	0.0011
Superlearner	0.0013	0.0011	0.0010	0.0013	0.0013	0.0013	0.0011	0.0013	0.0012
SVM	0.0013	0.0011	0.0012	0.0015	0.0015	0.0015	0.0014	0.0014	0.0014
Twang-GBM	0.0012	0.0011	0.0010	0.0013	0.0013	0.0012	0.0011	0.0013	0.0012
Simulated PS	0.0014	0.0014	0.0017	0.0019	0.0020	0.0018	0.0019	0.0022	0.0018

Table 20: (MCse; n=3000; ATE= -0.103) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0243	0.0262	0.0261	0.0284	0.0280	0.0279	0.0261	0.0267	0.0267
BAG-CART	0.0558	0.0869	0.0799	0.0783	0.0881	0.0798	0.0847	0.0826	0.0795
BOOSTLR	0.0430	0.0402	0.0470	0.0534	0.0480	0.0610	0.0438	0.0445	0.0476
GBM	0.0245	0.0250	0.0261	0.0254	0.0268	0.0255	0.0265	0.0264	0.0258
GBM-Stack	0.0241	0.0248	0.0262	0.0265	0.0273	0.0272	0.0266	0.0262	0.0261
KNN	0.0708	0.0833	0.0670	0.0886	0.0936	0.0894	0.0714	0.0645	0.0786
LR	0.0251	0.0236	0.0233	0.0272	0.0253	0.0261	0.0231	0.0243	0.0248
LR-Stack	0.0253	0.0246	0.0267	0.0268	0.0269	0.0283	0.0274	0.0265	0.0266
NB	0.0298	0.0300	0.0257	0.0337	0.0317	0.0313	0.0271	0.0286	0.0297
NNET	0.0250	0.0275	0.0296	0.0318	0.0307	0.0313	0.0324	0.0316	0.0300
RF	0.0274	0.0273	0.0285	0.0291	0.0304	0.0289	0.0272	0.0299	0.0286
RLR	0.0249	0.0235	0.0232	0.0269	0.0251	0.0257	0.0231	0.0241	0.0246
SC	0.0221	0.0216	0.0219	0.0223	0.0222	0.0217	0.0220	0.0225	0.0220
Superlearner	0.0238	0.0239	0.0242	0.0252	0.0254	0.0250	0.0244	0.0248	0.0246
SVM	0.0253	0.0250	0.0264	0.0289	0.0283	0.0292	0.0271	0.0272	0.0272
Twang-GBM	0.0233	0.0234	0.0241	0.0243	0.0249	0.0238	0.0241	0.0239	0.0240
Simulated PS	0.0246	0.0291	0.0304	0.0315	0.0347	0.0310	0.0343	0.0323	0.0310

Table 21: (MCse; n=3000; ATE= -0.091) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0257	0.0271	0.0274	0.0296	0.0290	0.0292	0.0273	0.0283	0.0280
BAG-CART	0.0583	0.0892	0.0843	0.0797	0.0896	0.0837	0.0867	0.0853	0.0821
BOOSTLR	0.0458	0.0416	0.0495	0.0551	0.0497	0.0637	0.0449	0.0462	0.0495
GBM	0.0259	0.0260	0.0273	0.0266	0.0277	0.0267	0.0274	0.0278	0.0269
GBM-Stack	0.0255	0.0258	0.0275	0.0277	0.0282	0.0284	0.0276	0.0277	0.0273
KNN	0.0733	0.0853	0.0704	0.0906	0.0950	0.0906	0.0753	0.0676	0.0810
LR	0.0265	0.0248	0.0245	0.0286	0.0266	0.0274	0.0241	0.0259	0.0260
LR-Stack	0.0265	0.0256	0.0279	0.0280	0.0278	0.0294	0.0284	0.0279	0.0277
NB	0.0318	0.0309	0.0269	0.0355	0.0333	0.0329	0.0279	0.0306	0.0312
NNET	0.0265	0.0284	0.0309	0.0326	0.0317	0.0327	0.0333	0.0331	0.0312
RF	0.0292	0.0287	0.0300	0.0304	0.0315	0.0312	0.0285	0.0321	0.0302
RLR	0.0263	0.0246	0.0244	0.0282	0.0264	0.0270	0.0240	0.0256	0.0258
SC	0.0233	0.0224	0.0228	0.0234	0.0233	0.0225	0.0229	0.0236	0.0230
Superlearner	0.0252	0.0249	0.0254	0.0264	0.0263	0.0262	0.0254	0.0262	0.0257
SVM	0.0266	0.0260	0.0276	0.0300	0.0291	0.0303	0.0282	0.0285	0.0283
Twang-GBM	0.0247	0.0244	0.0252	0.0254	0.0259	0.0249	0.0250	0.0253	0.0251
Simulated PS	0.0262	0.0299	0.0314	0.0325	0.0354	0.0323	0.0350	0.0339	0.0321

Table 22: (MCse; n=3000; ATE= -0.078) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0273	0.0282	0.0283	0.0309	0.0299	0.0308	0.0281	0.0300	0.0292
BAG-CART	0.0625	0.0926	0.0858	0.0832	0.0927	0.0867	0.0912	0.0889	0.0854
BOOSTLR	0.0481	0.0431	0.0516	0.0577	0.0523	0.0654	0.0479	0.0480	0.0518
GBM	0.0275	0.0270	0.0280	0.0277	0.0286	0.0279	0.0281	0.0290	0.0280
GBM-Stack	0.0270	0.0269	0.0281	0.0290	0.0292	0.0295	0.0284	0.0290	0.0284
KNN	0.0762	0.0886	0.0740	0.0921	0.0965	0.0924	0.0781	0.0712	0.0836
LR	0.0281	0.0260	0.0254	0.0301	0.0278	0.0291	0.0248	0.0271	0.0273
LR-Stack	0.0277	0.0268	0.0284	0.0291	0.0288	0.0303	0.0292	0.0291	0.0287
NB	0.0343	0.0327	0.0276	0.0379	0.0342	0.0347	0.0293	0.0327	0.0329
NNET	0.0281	0.0295	0.0321	0.0346	0.0326	0.0340	0.0346	0.0344	0.0325
RF	0.0315	0.0305	0.0309	0.0320	0.0329	0.0324	0.0293	0.0342	0.0317
RLR	0.0279	0.0259	0.0253	0.0296	0.0276	0.0286	0.0247	0.0268	0.0270
SC	0.0245	0.0235	0.0236	0.0242	0.0241	0.0236	0.0234	0.0245	0.0239
Superlearner	0.0267	0.0261	0.0260	0.0276	0.0273	0.0272	0.0260	0.0273	0.0268
SVM	0.0280	0.0272	0.0284	0.0311	0.0301	0.0312	0.0290	0.0295	0.0293
Twang-GBM	0.0262	0.0254	0.0258	0.0265	0.0268	0.0260	0.0256	0.0264	0.0261
Simulated PS	0.0277	0.0309	0.0321	0.0338	0.0362	0.0334	0.0360	0.0355	0.0332

Table 23: (MCse; n=3000; ATE= -0.062) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0288	0.0297	0.0289	0.0323	0.0314	0.0320	0.0292	0.0318	0.0305
BAG-CART	0.0670	0.0959	0.0886	0.0872	0.0969	0.0889	0.0951	0.0920	0.0889
BOOSTLR	0.0499	0.0452	0.0535	0.0609	0.0548	0.0673	0.0511	0.0504	0.0542
GBM	0.0291	0.0285	0.0284	0.0289	0.0301	0.0289	0.0291	0.0308	0.0292
GBM-Stack	0.0285	0.0284	0.0286	0.0302	0.0306	0.0305	0.0295	0.0306	0.0296
KNN	0.0782	0.0911	0.0766	0.0950	0.1002	0.0954	0.0793	0.0746	0.0863
LR	0.0297	0.0274	0.0259	0.0319	0.0296	0.0300	0.0260	0.0288	0.0287
LR-Stack	0.0292	0.0283	0.0288	0.0302	0.0302	0.0313	0.0303	0.0307	0.0299
NB	0.0362	0.0343	0.0279	0.0407	0.0360	0.0356	0.0301	0.0351	0.0345
NNET	0.0297	0.0310	0.0334	0.0362	0.0342	0.0352	0.0360	0.0367	0.0341
RF	0.0335	0.0328	0.0317	0.0336	0.0355	0.0340	0.0307	0.0369	0.0336
RLR	0.0295	0.0272	0.0257	0.0314	0.0294	0.0295	0.0260	0.0285	0.0284
SC	0.0258	0.0246	0.0237	0.0254	0.0252	0.0243	0.0244	0.0259	0.0249
Superlearner	0.0282	0.0276	0.0264	0.0288	0.0287	0.0283	0.0271	0.0290	0.0280
SVM	0.0292	0.0287	0.0288	0.0323	0.0314	0.0321	0.0300	0.0311	0.0304
Twang-GBM	0.0276	0.0269	0.0262	0.0277	0.0281	0.0270	0.0266	0.0280	0.0273
Simulated PS	0.0292	0.0323	0.0331	0.0352	0.0378	0.0345	0.0373	0.0376	0.0346

Table 24: (MCse; n=3000; ATE= -0.044) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0296	0.0306	0.0302	0.0337	0.0328	0.0338	0.0310	0.0331	0.0319
BAG-CART	0.0711	0.0988	0.0916	0.0904	0.1000	0.0924	0.0989	0.0954	0.0923
BOOSTLR	0.0516	0.0478	0.0566	0.0642	0.0571	0.0695	0.0540	0.0535	0.0568
GBM	0.0298	0.0295	0.0296	0.0306	0.0316	0.0303	0.0308	0.0320	0.0305
GBM-Stack	0.0294	0.0293	0.0300	0.0318	0.0319	0.0319	0.0311	0.0318	0.0309
KNN	0.0804	0.0946	0.0802	0.0978	0.1031	0.0978	0.0811	0.0783	0.0892
LR	0.0305	0.0287	0.0270	0.0342	0.0316	0.0315	0.0275	0.0300	0.0301
LR-Stack	0.0300	0.0292	0.0302	0.0318	0.0315	0.0329	0.0320	0.0318	0.0312
NB	0.0372	0.0356	0.0293	0.0441	0.0377	0.0375	0.0322	0.0370	0.0363
NNET	0.0305	0.0319	0.0347	0.0381	0.0357	0.0374	0.0379	0.0381	0.0355
RF	0.0345	0.0342	0.0333	0.0360	0.0376	0.0362	0.0330	0.0387	0.0354
RLR	0.0303	0.0285	0.0269	0.0336	0.0313	0.0310	0.0274	0.0297	0.0298
SC	0.0266	0.0253	0.0245	0.0267	0.0262	0.0255	0.0258	0.0267	0.0259
Superlearner	0.0290	0.0285	0.0277	0.0302	0.0301	0.0297	0.0288	0.0301	0.0293
SVM	0.0300	0.0293	0.0302	0.0337	0.0327	0.0335	0.0318	0.0321	0.0317
Twang-GBM	0.0284	0.0278	0.0275	0.0293	0.0296	0.0284	0.0282	0.0289	0.0285
Simulated PS	0.0300	0.0331	0.0348	0.0366	0.0393	0.0362	0.0388	0.0387	0.0359

Table 25: (MCse; n=3000; ATE= 0.055) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0361	0.0354	0.0351	0.0413	0.0394	0.0407	0.0362	0.0403	0.0381
BAG-CART	0.0943	0.1090	0.1066	0.1084	0.1197	0.1108	0.1126	0.1145	0.1095
BOOSTLR	0.0606	0.0558	0.0696	0.0769	0.0700	0.0811	0.0639	0.0669	0.0681
GBM	0.0365	0.0341	0.0341	0.0368	0.0374	0.0357	0.0353	0.0382	0.0360
GBM-Stack	0.0356	0.0338	0.0343	0.0374	0.0377	0.0367	0.0362	0.0378	0.0362
KNN	0.0948	0.1079	0.0943	0.1110	0.1132	0.1102	0.0961	0.0917	0.1024
LR	0.0378	0.0344	0.0312	0.0439	0.0400	0.0386	0.0316	0.0373	0.0369
LR-Stack	0.0358	0.0339	0.0348	0.0376	0.0375	0.0382	0.0370	0.0378	0.0366
NB	0.0473	0.0410	0.0337	0.0587	0.0475	0.0469	0.0371	0.0451	0.0447
NNET	0.0373	0.0374	0.0414	0.0459	0.0418	0.0455	0.0441	0.0460	0.0424
RF	0.0459	0.0422	0.0402	0.0472	0.0461	0.0452	0.0398	0.0479	0.0443
RLR	0.0374	0.0341	0.0311	0.0430	0.0393	0.0378	0.0316	0.0368	0.0364
SC	0.0317	0.0295	0.0282	0.0314	0.0310	0.0298	0.0297	0.0319	0.0304
Superlearner	0.0353	0.0333	0.0318	0.0364	0.0361	0.0350	0.0334	0.0363	0.0347
SVM	0.0354	0.0336	0.0348	0.0386	0.0383	0.0379	0.0364	0.0377	0.0366
Twang-GBM	0.0342	0.0323	0.0315	0.0349	0.0353	0.0335	0.0327	0.0346	0.0336
Simulated PS	0.0370	0.0375	0.0414	0.0438	0.0450	0.0424	0.0437	0.0467	0.0422

Table 26: (SE; n=3000; ATE= -0.103) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0243	0.0263	0.0256	0.0277	0.0270	0.0271	0.0259	0.0260	0.0262
BAG-CART	0.0391	0.0607	0.0574	0.0549	0.0608	0.0557	0.0577	0.0554	0.0552
BOOSTLR	0.0364	0.0360	0.0381	0.0416	0.0386	0.0458	0.0361	0.0358	0.0386
GBM	0.0243	0.0254	0.0253	0.0254	0.0257	0.0251	0.0258	0.0258	0.0254
GBM-Stack	0.0243	0.0252	0.0256	0.0264	0.0264	0.0263	0.0261	0.0258	0.0258
KNN	0.0573	0.0627	0.0521	0.0694	0.0696	0.0729	0.0584	0.0531	0.0619
LR	0.0249	0.0237	0.0234	0.0259	0.0244	0.0253	0.0233	0.0235	0.0243
LR-Stack	0.0246	0.0250	0.0260	0.0263	0.0260	0.0268	0.0267	0.0262	0.0259
NB	0.0279	0.0284	0.0244	0.0300	0.0287	0.0290	0.0252	0.0265	0.0275
NNET	0.0247	0.0273	0.0277	0.0296	0.0289	0.0294	0.0287	0.0288	0.0281
RF	0.0260	0.0267	0.0264	0.0281	0.0281	0.0274	0.0264	0.0273	0.0271
RLR	0.0247	0.0236	0.0233	0.0256	0.0242	0.0250	0.0232	0.0234	0.0241
SC	0.0224	0.0219	0.0219	0.0223	0.0219	0.0221	0.0220	0.0221	0.0221
Superlearner	0.0240	0.0243	0.0238	0.0250	0.0247	0.0248	0.0242	0.0244	0.0244
SVM	0.0255	0.0255	0.0257	0.0280	0.0271	0.0278	0.0263	0.0265	0.0266
Twang-GBM	0.0238	0.0240	0.0238	0.0246	0.0243	0.0242	0.0239	0.0239	0.0241
Simulated PS	0.0247	0.0277	0.0279	0.0293	0.0303	0.0293	0.0294	0.0293	0.0285

Table 27: (SE; n=3000; ATE= -0.091) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0254	0.0272	0.0266	0.0288	0.0279	0.0281	0.0270	0.0273	0.0273
BAG-CART	0.0415	0.0626	0.0608	0.0575	0.0633	0.0581	0.0600	0.0579	0.0577
BOOSTLR	0.0381	0.0372	0.0402	0.0434	0.0402	0.0477	0.0376	0.0372	0.0402
GBM	0.0253	0.0263	0.0262	0.0265	0.0267	0.0260	0.0268	0.0269	0.0263
GBM-Stack	0.0253	0.0261	0.0265	0.0275	0.0273	0.0272	0.0271	0.0269	0.0267
KNN	0.0601	0.0653	0.0559	0.0718	0.0720	0.0752	0.0621	0.0564	0.0648
LR	0.0260	0.0246	0.0243	0.0272	0.0255	0.0263	0.0241	0.0246	0.0253
LR-Stack	0.0257	0.0259	0.0270	0.0273	0.0269	0.0277	0.0276	0.0272	0.0269
NB	0.0293	0.0294	0.0253	0.0317	0.0300	0.0301	0.0261	0.0279	0.0287
NNET	0.0259	0.0282	0.0289	0.0308	0.0299	0.0305	0.0299	0.0301	0.0293
RF	0.0274	0.0279	0.0278	0.0295	0.0294	0.0290	0.0278	0.0290	0.0285
RLR	0.0259	0.0245	0.0242	0.0269	0.0253	0.0261	0.0240	0.0244	0.0252
SC	0.0233	0.0227	0.0227	0.0233	0.0227	0.0229	0.0227	0.0230	0.0229
Superlearner	0.0250	0.0251	0.0247	0.0261	0.0257	0.0257	0.0252	0.0255	0.0254
SVM	0.0265	0.0263	0.0267	0.0290	0.0280	0.0287	0.0273	0.0274	0.0275
Twang-GBM	0.0248	0.0249	0.0247	0.0257	0.0253	0.0251	0.0248	0.0250	0.0250
Simulated PS	0.0258	0.0286	0.0290	0.0304	0.0312	0.0304	0.0305	0.0305	0.0296

Table 28: (SE; n=3000; ATE= -0.078) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0266	0.0282	0.0278	0.0303	0.0290	0.0294	0.0281	0.0288	0.0285
BAG-CART	0.0443	0.0654	0.0635	0.0614	0.0661	0.0604	0.0633	0.0613	0.0607
BOOSTLR	0.0398	0.0386	0.0421	0.0455	0.0420	0.0496	0.0396	0.0389	0.0420
GBM	0.0265	0.0272	0.0273	0.0277	0.0277	0.0271	0.0278	0.0282	0.0274
GBM-Stack	0.0264	0.0270	0.0276	0.0287	0.0283	0.0283	0.0282	0.0281	0.0278
KNN	0.0633	0.0686	0.0596	0.0745	0.0745	0.0780	0.0656	0.0598	0.0680
LR	0.0273	0.0257	0.0253	0.0287	0.0268	0.0275	0.0251	0.0258	0.0265
LR-Stack	0.0268	0.0268	0.0281	0.0285	0.0279	0.0287	0.0287	0.0284	0.0280
NB	0.0309	0.0305	0.0264	0.0336	0.0315	0.0314	0.0273	0.0298	0.0302
NNET	0.0271	0.0292	0.0301	0.0325	0.0309	0.0317	0.0312	0.0316	0.0305
RF	0.0289	0.0294	0.0292	0.0312	0.0308	0.0303	0.0291	0.0311	0.0300
RLR	0.0271	0.0255	0.0252	0.0284	0.0266	0.0272	0.0250	0.0257	0.0263
SC	0.0243	0.0235	0.0236	0.0243	0.0236	0.0238	0.0235	0.0240	0.0238
Superlearner	0.0261	0.0261	0.0258	0.0274	0.0267	0.0268	0.0262	0.0267	0.0265
SVM	0.0275	0.0272	0.0278	0.0301	0.0289	0.0296	0.0284	0.0286	0.0285
Twang-GBM	0.0259	0.0258	0.0258	0.0269	0.0263	0.0262	0.0258	0.0262	0.0261
Simulated PS	0.0270	0.0296	0.0302	0.0318	0.0323	0.0316	0.0318	0.0320	0.0308

Table 29: (SE; n=3000; ATE= -0.062) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0279	0.0292	0.0290	0.0318	0.0302	0.0307	0.0292	0.0303	0.0298
BAG-CART	0.0481	0.0685	0.0668	0.0655	0.0699	0.0632	0.0660	0.0648	0.0641
BOOSTLR	0.0416	0.0402	0.0444	0.0477	0.0438	0.0518	0.0418	0.0406	0.0440
GBM	0.0277	0.0282	0.0285	0.0291	0.0289	0.0282	0.0289	0.0296	0.0286
GBM-Stack	0.0277	0.0281	0.0288	0.0300	0.0294	0.0294	0.0293	0.0295	0.0290
KNN	0.0666	0.0717	0.0638	0.0778	0.0774	0.0810	0.0696	0.0636	0.0714
LR	0.0286	0.0267	0.0264	0.0305	0.0282	0.0288	0.0261	0.0272	0.0278
LR-Stack	0.0280	0.0278	0.0292	0.0298	0.0290	0.0298	0.0298	0.0297	0.0291
NB	0.0325	0.0316	0.0275	0.0360	0.0330	0.0328	0.0284	0.0317	0.0317
NNET	0.0284	0.0303	0.0317	0.0340	0.0323	0.0331	0.0326	0.0333	0.0319
RF	0.0307	0.0309	0.0308	0.0331	0.0326	0.0319	0.0305	0.0330	0.0317
RLR	0.0284	0.0266	0.0263	0.0302	0.0280	0.0284	0.0260	0.0270	0.0276
SC	0.0254	0.0245	0.0245	0.0254	0.0246	0.0248	0.0244	0.0252	0.0248
Superlearner	0.0273	0.0271	0.0269	0.0287	0.0278	0.0279	0.0272	0.0281	0.0276
SVM	0.0286	0.0282	0.0290	0.0313	0.0300	0.0307	0.0295	0.0299	0.0296
Twang-GBM	0.0271	0.0269	0.0269	0.0282	0.0275	0.0273	0.0268	0.0275	0.0273
Simulated PS	0.0283	0.0306	0.0316	0.0333	0.0335	0.0329	0.0330	0.0336	0.0321

Table 30: (SE; n=3000; ATE= -0.044) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0293	0.0303	0.0304	0.0334	0.0315	0.0322	0.0305	0.0320	0.0312
BAG-CART	0.0533	0.0717	0.0701	0.0696	0.0738	0.0674	0.0700	0.0693	0.0681
BOOSTLR	0.0435	0.0420	0.0470	0.0503	0.0459	0.0539	0.0442	0.0427	0.0462
GBM	0.0291	0.0293	0.0297	0.0307	0.0302	0.0295	0.0301	0.0310	0.0300
GBM-Stack	0.0291	0.0292	0.0301	0.0315	0.0306	0.0307	0.0306	0.0309	0.0303
KNN	0.0706	0.0749	0.0680	0.0816	0.0807	0.0841	0.0740	0.0677	0.0752
LR	0.0301	0.0279	0.0276	0.0326	0.0300	0.0302	0.0272	0.0287	0.0293
LR-Stack	0.0294	0.0289	0.0305	0.0312	0.0302	0.0310	0.0310	0.0311	0.0304
NB	0.0345	0.0330	0.0288	0.0388	0.0348	0.0344	0.0298	0.0336	0.0335
NNET	0.0299	0.0315	0.0334	0.0358	0.0337	0.0349	0.0339	0.0351	0.0335
RF	0.0327	0.0324	0.0324	0.0355	0.0344	0.0337	0.0321	0.0351	0.0335
RLR	0.0299	0.0278	0.0275	0.0322	0.0297	0.0298	0.0270	0.0286	0.0291
SC	0.0266	0.0254	0.0255	0.0266	0.0257	0.0259	0.0253	0.0264	0.0259
Superlearner	0.0287	0.0282	0.0281	0.0302	0.0291	0.0292	0.0284	0.0296	0.0289
SVM	0.0299	0.0292	0.0302	0.0327	0.0312	0.0319	0.0308	0.0312	0.0309
Twang-GBM	0.0285	0.0280	0.0281	0.0297	0.0287	0.0286	0.0280	0.0289	0.0285
Simulated PS	0.0299	0.0318	0.0331	0.0350	0.0348	0.0345	0.0344	0.0352	0.0336

Table 31: (SE; n=3000; ATE= 0.055) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=3000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0356	0.0355	0.0363	0.0409	0.0374	0.0396	0.0363	0.0392	0.0376
BAG-CART	0.0746	0.0877	0.0879	0.0906	0.0927	0.0875	0.0883	0.0886	0.0872
BOOSTLR	0.0521	0.0495	0.0582	0.0616	0.0557	0.0661	0.0529	0.0521	0.0560
GBM	0.0353	0.0343	0.0351	0.0372	0.0360	0.0354	0.0353	0.0372	0.0357
GBM-Stack	0.0351	0.0341	0.0356	0.0378	0.0361	0.0364	0.0359	0.0371	0.0360
KNN	0.0880	0.0923	0.0861	0.0963	0.0959	0.0996	0.0919	0.0845	0.0918
LR	0.0370	0.0336	0.0325	0.0420	0.0379	0.0376	0.0316	0.0359	0.0360
LR-Stack	0.0354	0.0339	0.0360	0.0374	0.0357	0.0366	0.0362	0.0371	0.0361
NB	0.0437	0.0389	0.0342	0.0524	0.0437	0.0438	0.0350	0.0427	0.0418
NNET	0.0366	0.0370	0.0402	0.0437	0.0399	0.0433	0.0406	0.0434	0.0406
RF	0.0420	0.0398	0.0396	0.0451	0.0430	0.0431	0.0393	0.0447	0.0421
RLR	0.0366	0.0334	0.0323	0.0413	0.0373	0.0370	0.0315	0.0355	0.0356
SC	0.0317	0.0298	0.0298	0.0319	0.0304	0.0305	0.0295	0.0315	0.0306
Superlearner	0.0347	0.0332	0.0332	0.0367	0.0348	0.0350	0.0334	0.0358	0.0346
SVM	0.0351	0.0339	0.0357	0.0385	0.0365	0.0371	0.0361	0.0369	0.0362
Twang-GBM	0.0343	0.0329	0.0332	0.0359	0.0342	0.0342	0.0328	0.0347	0.0340
Simulated PS	0.0366	0.0369	0.0398	0.0424	0.0406	0.0412	0.0411	0.0430	0.0402

Table 32: (Raw (naive) bias; n=3000) Summary table of relative ATE estimation bias (in %) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Relative bias is presented across all eight simulated scenarios (A-H) and across all six simulated average treatment effects based on the n=3000 data setup.

	A	B	C	D	E	F	G	H	Mean-bias
-1.2	18.31	19.11	4.43	23.29	22.78	20.85	5.33	19.67	16.72
-1.0	22.02	22.71	5.35	28.50	27.56	25.30	6.55	23.76	20.22
-0.8	27.60	28.30	6.93	36.04	34.46	31.74	8.47	29.98	25.44
-0.6	37.99	37.62	9.72	49.04	46.58	42.46	11.07	41.56	34.51
-0.4	57.95	56.70	14.54	75.27	69.80	65.14	16.82	64.35	52.57
+0.4	62.87	61.27	12.91	81.27	74.23	67.56	16.31	68.80	55.65

Table 33: (n=3000) Summary table of ATE estimation mean-squared error (mse) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. mean-squared error (mse) computed across m=1000 simulated data set based on the n=3000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-mse
-1.2	0.0008	0.0009	0.0005	0.0011	0.0011	0.0009	0.0005	0.0009	0.0008
-1.0	0.0009	0.0009	0.0005	0.0012	0.0012	0.0010	0.0005	0.0010	0.0009
-0.8	0.0011	0.0010	0.0006	0.0014	0.0013	0.0012	0.0006	0.0012	0.0010
-0.6	0.0012	0.0012	0.0006	0.0016	0.0015	0.0013	0.0006	0.0014	0.0012
-0.4	0.0014	0.0013	0.0006	0.0018	0.0017	0.0015	0.0007	0.0016	0.0013
+0.4	0.0022	0.0021	0.0008	0.0031	0.0027	0.0023	0.0010	0.0025	0.0021

Table 34: (n=3000) Summary table of ATE estimation Monte Carlo standard error (MCse) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Monte Carlo standard error (MCse) computed across m=1000 simulated data set based on the n=3000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-MCse
-1.2	0.0220	0.0217	0.0217	0.0222	0.0226	0.0215	0.0218	0.0230	0.0221
-1.0	0.0230	0.0226	0.0227	0.0234	0.0237	0.0223	0.0226	0.0240	0.0230
-0.8	0.0242	0.0235	0.0234	0.0244	0.0247	0.0234	0.0231	0.0249	0.0240
-0.6	0.0255	0.0247	0.0235	0.0258	0.0256	0.0242	0.0242	0.0264	0.0250
-0.4	0.0265	0.0253	0.0243	0.0269	0.0267	0.0255	0.0256	0.0272	0.0260
+0.4	0.0314	0.0294	0.0278	0.0313	0.0308	0.0299	0.0295	0.0321	0.0303

Table 35: (n=3000) Summary table of ATE estimation Robust sandwich-type standard error (SE) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Robust sandwich-type standard error (SE) computed across m=1000 simulated data set based on the n=3000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-bias
-1.2	0.0224	0.0219	0.0218	0.0225	0.0221	0.0221	0.0218	0.0225	0.0221
-1.0	0.0233	0.0227	0.0225	0.0235	0.0230	0.0230	0.0225	0.0235	0.0230
-0.8	0.0244	0.0236	0.0234	0.0246	0.0239	0.0239	0.0233	0.0245	0.0240
-0.6	0.0255	0.0246	0.0243	0.0257	0.0249	0.0249	0.0242	0.0257	0.0250
-0.4	0.0267	0.0256	0.0252	0.0269	0.0260	0.0260	0.0251	0.0269	0.0261
+0.4	0.0316	0.0299	0.0295	0.0318	0.0305	0.0305	0.0292	0.0318	0.0306

Table 36: (ASAM; n=3000) Mean of the averaged standardized absolute mean differences over m=1000 simulated data sets in the n=3000 simulated data setup. In each data set the mean of the standardized absolute mean differences of all ten covariates is taken. We describe values in this table as ASAM. The last row (NO WEIGHT) presents the ASAM in the initial non-weighted data.

	A	B	C	D	E	F	G	H	Mean-ASAM
AVNNET	0.088	0.085	0.085	0.102	0.090	0.098	0.084	0.095	0.091
BAG-CART	0.248	0.271	0.261	0.288	0.292	0.291	0.255	0.258	0.270
BOOSTLR	0.298	0.274	0.275	0.359	0.359	0.396	0.257	0.319	0.317
GBM	0.092	0.090	0.090	0.098	0.093	0.097	0.084	0.092	0.092
GBM-Stack	0.087	0.084	0.088	0.094	0.088	0.090	0.085	0.090	0.088
KNN	0.336	0.373	0.252	0.384	0.411	0.411	0.292	0.321	0.347
LR	0.088	0.092	0.079	0.137	0.125	0.097	0.074	0.110	0.100
LR-Stack	0.090	0.088	0.097	0.100	0.090	0.096	0.093	0.094	0.093
NB	0.196	0.149	0.096	0.348	0.225	0.209	0.098	0.177	0.187
NNET	0.087	0.089	0.097	0.111	0.096	0.112	0.098	0.111	0.100
RF	0.105	0.100	0.095	0.116	0.109	0.108	0.093	0.109	0.104
RLR	0.087	0.090	0.078	0.130	0.119	0.094	0.073	0.106	0.097
SC	0.184	0.161	0.111	0.201	0.180	0.207	0.118	0.145	0.164
Superlearner	0.088	0.084	0.085	0.094	0.088	0.094	0.078	0.089	0.087
SVM	0.103	0.089	0.087	0.109	0.094	0.099	0.088	0.096	0.096
Twang-GBM	0.091	0.090	0.079	0.098	0.097	0.098	0.073	0.092	0.090
Simulated PS	0.076	0.078	0.089	0.093	0.087	0.091	0.094	0.096	0.088
NO WEIGHT	0.287	0.264	0.147	0.338	0.310	0.322	0.153	0.254	0.259

Table 37: (ASAM$_{conf}$; n=3000) Mean of the averaged standardized absolute mean differences over m=1000 simulated data sets in the n=3000 simulated data setup. In each data set, the mean of the standardized absolute mean differences of the four confounding covariates is taken. We therefore describe values in this table with ASAM$_{conf}$. The last row (NO WEIGHT) presents the ASAM$_{conf}$ in the initial non-weighted data.

	A	B	C	D	E	F	G	H	Mean-ASAM
AVN	0.089	0.085	0.086	0.101	0.089	0.099	0.085	0.097	0.091
BAG-CART	0.252	0.282	0.257	0.298	0.302	0.321	0.249	0.273	0.279
BOOSTLR	0.288	0.256	0.327	0.454	0.458	0.532	0.299	0.385	0.375
GBM	0.096	0.092	0.095	0.105	0.097	0.109	0.088	0.094	0.097
GBM-Stack	0.087	0.082	0.088	0.094	0.088	0.094	0.088	0.093	0.089
KNN	0.326	0.376	0.258	0.393	0.447	0.446	0.305	0.332	0.360
LR	0.088	0.101	0.075	0.150	0.145	0.102	0.075	0.129	0.108
LR-Stack	0.089	0.085	0.095	0.103	0.089	0.101	0.098	0.097	0.095
NB	0.203	0.147	0.091	0.386	0.239	0.217	0.095	0.189	0.196
NNET	0.087	0.089	0.100	0.112	0.095	0.115	0.101	0.115	0.102
RF	0.102	0.096	0.102	0.108	0.102	0.103	0.095	0.108	0.102
RLR	0.088	0.098	0.075	0.142	0.137	0.099	0.074	0.124	0.105
SC	0.199	0.167	0.127	0.253	0.225	0.305	0.150	0.185	0.201
Superlearner	0.090	0.084	0.085	0.096	0.089	0.105	0.083	0.092	0.091
SVM	0.106	0.087	0.088	0.106	0.091	0.107	0.093	0.097	0.097
Twang-GBM	0.097	0.093	0.083	0.107	0.106	0.119	0.083	0.102	0.098
Simulated PS	0.077	0.079	0.091	0.098	0.088	0.094	0.098	0.098	0.090
NO WEIGHT	0.316	0.289	0.160	0.424	0.389	0.459	0.201	0.337	0.322

Table 38: Average of maximum IPTW weights resulting through the different propensity score estimation models in each of the m=1000 simulated data sets for each scenario (A-H) in the n=3000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVNNET	15.4	15.6	19.8	27.7	17.8	30.2	20.6	25.4	21.6
BAG-CART	193.0	191.8	195.1	199.1	196.0	197.7	194.2	196.3	195.4
BOOSTLR	54.1	49.7	104.7	102.1	81.0	127.8	84.1	62.4	83.3
GBM	16.6	13.9	16.7	16.7	15.5	15.1	17.4	17.3	16.1
GBM-Stack	11.0	10.1	12.0	12.1	11.3	12.0	12.7	12.0	11.7
KNN	200.0	200.0	200.0	200.0	200.0	200.0	199.9	200.0	200.0
LR	23.5	22.0	10.2	55.5	42.1	32.0	7.1	29.9	27.8
LR-Stack	13.1	9.6	14.1	13.2	10.3	13.2	14.0	11.7	12.4
NB	58.6	34.3	18.3	143.6	62.4	61.6	26.1	102.7	63.5
NNET	19.5	22.2	33.7	39.1	26.1	43.7	36.6	43.0	33.0
RF	55.9	42.2	38.4	63.7	51.4	58.2	36.8	63.2	51.2
RLR	22.2	20.7	9.9	51.1	39.0	29.3	6.8	27.8	25.8
SC	6.4	5.7	3.8	8.7	7.5	6.0	3.2	7.5	6.1
Superlearner	11.4	9.8	10.7	12.8	10.9	12.1	10.8	10.2	11.1
SVM	12.5	12.2	17.2	18.3	16.7	18.5	16.9	16.6	16.1
Twang-GBM	13.1	11.1	11.3	15.8	13.0	13.9	12.0	14.9	13.1
Simulated PS	22.0	22.8	36.2	32.3	28.8	33.2	42.0	37.9	31.9

Table 39: Average of mean IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=3000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVNNET	2.01	2.02	1.99	2.05	2.02	2.04	1.97	2.03	2.02
BAG-CART	3.21	3.18	3.29	3.66	3.44	3.49	3.28	3.37	3.37
BOOSTLR	3.02	2.99	3.08	3.52	3.37	3.68	3.00	3.09	3.22
GBM	1.99	1.98	1.95	1.97	1.98	1.95	1.96	2.00	1.97
GBM-Stack	2.00	2.00	1.99	2.00	2.00	2.00	2.00	2.00	2.00
KNN	6.29	5.78	5.50	6.64	6.56	6.18	5.54	6.46	6.12
LR	2.02	2.05	2.03	2.13	2.13	2.04	2.01	2.09	2.06
LR-Stack	2.01	2.01	2.01	2.02	2.00	2.04	2.02	2.00	2.01
NB	2.46	2.31	1.93	2.96	2.58	2.34	2.05	3.52	2.52
NNET	2.02	2.06	2.08	2.11	2.07	2.13	2.06	2.13	2.08
RF	2.34	2.25	2.22	2.31	2.27	2.26	2.16	2.32	2.27
RLR	2.00	2.04	2.02	2.11	2.10	2.02	2.01	2.07	2.05
SC	1.90	1.92	1.97	1.88	1.90	1.89	1.98	1.94	1.92
Superlearner	1.85	1.85	1.81	1.81	1.82	1.82	1.81	1.81	1.82
SVM	2.01	2.00	2.00	2.00	2.00	2.00	1.99	1.99	2.00
Twang-GBM	1.97	1.96	1.90	1.95	1.95	1.93	1.90	1.96	1.94
Simulated PS	2.00	2.00	2.00	2.00	2.00	2.00	1.99	1.99	2.00

Table 40: Average of the first quantile of the IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=3000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVN	1.20	1.23	1.22	1.14	1.18	1.16	1.22	1.17	1.19
BAG-CART	1.20	1.22	1.19	1.13	1.17	1.16	1.19	1.14	1.18
BOOSTLR	1.25	1.33	1.27	1.15	1.22	1.11	1.31	1.23	1.23
GBM	1.23	1.26	1.24	1.17	1.20	1.21	1.24	1.18	1.22
GBM-Stack	1.21	1.24	1.20	1.15	1.18	1.17	1.20	1.17	1.19
KNN	1.14	1.19	1.20	1.08	1.12	1.10	1.18	1.11	1.14
LR	1.21	1.31	1.45	1.17	1.23	1.21	1.48	1.25	1.29
LR-Stack	1.21	1.23	1.20	1.15	1.18	1.17	1.19	1.17	1.19
NB	1.13	1.17	1.27	1.10	1.13	1.16	1.26	1.06	1.16
NNET	1.20	1.23	1.20	1.13	1.17	1.15	1.19	1.15	1.18
RF	1.16	1.21	1.20	1.12	1.16	1.15	1.21	1.14	1.17
RLR	1.21	1.32	1.46	1.17	1.24	1.22	1.49	1.26	1.30
SC	1.34	1.51	1.57	1.32	1.42	1.42	1.63	1.36	1.45
SL	1.21	1.25	1.25	1.16	1.20	1.20	1.26	1.20	1.22
SVM	1.25	1.26	1.22	1.18	1.20	1.19	1.21	1.20	1.21
Twang-GBM	1.22	1.27	1.27	1.17	1.21	1.21	1.27	1.18	1.23
Simulated PS	1.21	1.23	1.20	1.13	1.17	1.15	1.18	1.13	1.18

Table 41: Average of the third quantile of the IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=3000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVN	2.11	2.15	2.10	1.92	2.01	2.00	2.07	1.96	2.04
BAG-CART	2.26	2.29	2.23	2.06	2.15	2.14	2.23	2.06	2.18
BOOSTLR	3.72	3.72	3.72	3.67	3.71	3.71	3.72	3.71	3.71
GBM	2.11	2.14	2.09	1.94	2.03	2.02	2.10	1.97	2.05
GBM-Stack	2.12	2.17	2.10	1.91	2.02	2.00	2.08	1.95	2.04
KNN	2.40	2.48	2.48	2.15	2.27	2.24	2.45	2.24	2.34
LR	2.09	2.16	2.24	1.96	2.05	2.09	2.28	2.08	2.12
LR-Stack	2.12	2.17	2.10	1.91	2.02	2.01	2.09	1.93	2.04
NB	2.18	2.23	2.10	2.02	2.10	2.15	2.23	2.30	2.16
NNET	2.09	2.16	2.11	1.92	2.02	2.02	2.09	1.98	2.05
RF	2.17	2.20	2.17	1.95	2.05	2.04	2.13	1.99	2.09
RLR	2.09	2.15	2.23	1.96	2.04	2.08	2.28	2.08	2.12
SC	2.27	2.12	2.39	2.13	2.09	2.15	2.37	2.32	2.23
SL	1.98	2.00	1.96	1.82	1.89	1.90	1.97	1.85	1.92
SVM	2.19	2.21	2.11	1.90	2.03	2.01	2.09	1.98	2.07
Twang-GBM	2.10	2.14	2.07	1.92	2.01	2.00	2.06	1.96	2.03
Simulated PS	2.08	2.11	2.04	1.88	1.97	1.94	2.02	1.86	1.99

12.2. Appendix A – Performance measure results based on n = 2000 setup

Table 42: (Relative Bias; n=2000; ATE= -0.103) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=2000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	1.80	1.18	1.09	0.81	0.86	0.33	0.89	0.90	0.98
BAG-CART	8.37	17.93	1.96	16.36	19.10	13.90	1.42	12.72	11.47
BOOSTLR	6.93	6.10	2.95	4.09	1.50	9.12	7.78	5.86	5.54
GBM	1.11	1.36	1.52	3.46	2.38	1.82	2.13	0.49	1.78
GBM-Stack	0.47	0.57	0.01	0.79	0.50	0.68	0.63	0.34	0.50
KNN	6.65	10.11	8.47	18.53	11.64	7.82	7.69	4.82	9.47
LR	0.51	2.15	2.47	0.16	0.25	0.39	0.92	1.06	0.99
LR-Stack	0.31	0.21	0.66	0.17	1.06	0.52	1.40	0.16	0.56
NB	14.83	10.12	1.68	19.78	13.36	15.96	3.56	6.35	10.71
NNET	1.15	0.15	0.49	0.83	1.31	2.38	0.26	2.03	1.08
RF	0.22	0.44	1.16	0.10	0.28	0.03	2.47	0.52	0.65
RLR	0.92	2.51	2.12	0.78	0.88	0.94	0.63	0.45	1.15
SC	8.02	9.95	3.27	11.53	10.79	8.61	5.83	8.51	8.31
Superlearner	0.46	0.50	0.53	1.90	2.09	1.77	1.65	2.16	1.38
SVM	2.04	0.59	0.95	0.38	0.01	2.38	1.03	0.35	0.97
Twang-GBM	4.57	5.53	1.44	7.58	6.79	5.68	2.96	5.73	5.04
Simulated PS	0.89	0.74	0.41	1.20	0.26	0.12	0.88	0.57	0.63

Table 43: (Relative Bias; n=2000; ATE= -0.091) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=2000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	2.10	1.27	1.40	1.14	0.58	0.43	0.81	0.82	1.07
BAG-CART	11.28	21.85	3.48	20.35	23.36	16.44	0.23	16.17	14.14
BOOSTLR	8.36	7.80	3.40	5.67	2.29	10.06	10.09	7.55	6.90
GBM	1.21	1.50	1.81	4.18	2.23	2.29	2.07	0.48	1.97
GBM-Stack	0.41	0.81	0.05	1.25	0.00	1.03	0.56	0.54	0.58
KNN	8.84	12.38	11.44	22.21	15.39	8.80	9.40	7.18	11.96
LR	0.45	1.54	3.41	0.78	0.88	0.14	1.84	2.18	1.40
LR-Stack	0.45	0.09	0.92	0.43	0.63	0.59	1.35	0.34	0.60
NB	18.21	12.55	1.90	25.46	17.29	19.61	3.73	9.42	13.52
NNET	1.29	0.34	0.93	0.82	2.08	2.85	0.65	2.46	1.43
RF	1.01	0.94	1.13	0.10	0.93	0.41	2.28	1.38	1.02
RLR	0.95	2.04	2.98	0.06	0.09	0.85	1.47	1.42	1.23
SC	9.52	11.51	3.68	13.66	12.19	10.29	6.59	9.83	9.66
Superlearner	0.74	0.29	0.58	2.05	1.73	2.13	1.66	2.19	1.42
SVM	2.59	0.50	1.23	0.93	0.70	3.06	1.42	0.53	1.37
Twang-GBM	5.33	6.36	1.69	8.99	7.43	6.77	3.11	6.66	5.79
Simulated PS	0.90	0.74	0.70	1.65	0.78	0.05	0.83	0.56	0.77

Table 44: (Relative Bias; n=2000; ATE= -0.078) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=2000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	2.07	1.25	1.50	1.66	0.72	0.12	0.85	0.11	1.04
BAG-CART	15.66	27.15	4.34	25.55	28.98	20.99	0.03	22.70	18.18
BOOSTLR	10.53	10.20	4.27	6.15	3.87	13.45	12.21	9.88	8.82
GBM	0.92	1.37	2.43	5.76	2.44	2.53	2.75	0.12	2.29
GBM-Stack	0.11	1.37	0.33	2.21	0.08	0.96	0.71	1.31	0.88
KNN	11.41	17.21	12.77	26.50	19.18	13.05	13.39	10.14	15.46
LR	0.20	0.81	4.63	1.58	2.72	0.51	3.09	3.94	2.19
LR-Stack	1.21	0.35	0.88	1.06	0.52	0.56	1.66	0.98	0.90
NB	23.97	16.49	2.53	33.48	22.88	25.46	4.84	14.40	18.01
NNET	0.96	0.90	1.11	0.31	2.56	3.90	0.50	4.20	1.81
RF	2.22	1.96	1.71	0.12	1.64	1.20	2.68	3.47	1.87
RLR	0.47	1.47	4.08	0.37	1.63	0.44	2.58	2.96	1.75
SC	11.31	13.87	4.51	17.69	14.69	12.62	7.99	11.86	11.82
Superlearner	1.51	0.20	0.90	3.00	1.70	2.32	1.98	1.75	1.67
SVM	3.10	0.27	1.37	2.13	0.92	3.69	1.84	0.19	1.69
Twang-GBM	6.21	7.45	2.21	11.76	9.04	8.13	3.84	7.65	7.04
Simulated PS	0.35	0.54	0.73	2.66	0.78	0.22	1.45	0.25	0.87

Table 45: (Relative Bias; n=2000; ATE= -0.062) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=2000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	2.69	2.85	1.64	1.61	1.10	0.35	0.66	0.16	1.38
BAG-CART	20.52	33.57	7.75	37.04	37.85	27.44	0.30	30.17	24.33
BOOSTLR	12.63	12.25	4.15	10.82	5.54	17.95	15.17	13.67	11.52
GBM	1.26	2.81	3.50	7.18	3.13	4.20	3.28	0.18	3.19
GBM-Stack	0.14	0.83	0.64	2.61	0.11	1.74	0.40	1.04	0.94
KNN	14.53	19.57	16.50	34.22	24.55	18.04	20.39	12.96	20.10
LR	0.29	1.18	6.51	4.14	5.36	0.65	4.96	6.31	3.67
LR-Stack	1.56	0.46	0.85	1.04	0.61	1.34	1.55	0.67	1.01
NB	32.20	21.20	3.49	47.30	32.57	34.39	5.89	20.91	24.74
NNET	1.10	0.36	1.06	0.69	3.00	5.01	1.21	4.81	2.15
RF	3.65	1.70	2.51	0.95	2.75	1.66	2.80	4.89	2.61
RLR	0.59	2.08	5.70	2.37	3.76	0.69	4.23	4.93	3.04
SC	14.71	18.82	6.12	23.08	19.49	17.44	10.31	16.00	15.75
Superlearner	1.94	0.62	1.47	3.26	1.89	3.48	2.20	2.40	2.16
SVM	4.65	1.13	1.77	2.90	1.21	5.74	2.91	0.94	2.66
Twang-GBM	8.20	10.60	3.32	14.95	11.87	11.37	4.69	10.62	9.45
Simulated PS	0.37	1.98	0.81	3.23	0.81	0.11	1.80	0.24	1.17

Table 46: (Relative Bias; n=2000; ATE= -0.044) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=2000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	3.88	5.10	2.83	0.26	2.27	0.23	0.19	0.95	1.96
BAG-CART	35.65	50.63	7.66	59.16	54.75	42.51	1.84	48.72	37.61
BOOSTLR	19.68	18.66	6.05	21.89	7.97	27.31	21.89	20.74	18.03
GBM	1.92	4.89	5.17	8.71	4.83	6.66	4.04	0.62	4.61
GBM-Stack	0.16	0.29	0.30	2.14	0.02	3.12	0.05	2.48	1.07
KNN	24.74	31.44	23.46	54.34	38.14	31.60	28.71	22.75	31.90
LR	1.15	1.04	9.81	11.23	8.81	1.94	8.50	12.58	6.88
LR-Stack	2.53	1.54	1.79	0.51	0.93	2.36	1.60	1.31	1.57
NB	50.79	31.88	5.11	79.42	51.20	53.21	8.50	36.85	39.62
NNET	1.27	1.24	2.18	2.74	4.38	8.54	2.98	7.84	3.90
RF	5.86	2.98	2.84	4.49	5.20	2.80	3.59	9.08	4.60
RLR	0.25	2.52	8.62	8.32	6.41	0.22	7.37	10.31	5.50
SC	22.47	28.37	9.40	33.31	29.42	26.62	15.34	23.08	23.50
Superlearner	3.24	1.32	1.93	2.28	2.60	5.26	2.74	2.57	2.74
SVM	7.96	2.39	3.21	2.92	1.55	9.33	4.83	1.32	4.19
Twang-GBM	12.42	16.46	4.67	20.64	17.93	17.16	6.50	15.27	13.88
Simulated PS	0.02	3.90	1.21	3.24	0.41	0.88	1.64	0.08	1.42

Table 47: (Relative Bias; n=2000; ATE= 0.055) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=2000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	3.96	3.73	1.40	5.39	1.43	3.31	2.20	6.07	3.44
BAG-CART	40.10	51.28	17.05	58.67	55.34	49.11	3.31	54.84	41.21
BOOSTLR	20.35	19.17	4.22	29.88	15.49	35.20	24.32	27.89	22.07
GBM	1.77	3.26	6.91	4.69	0.10	5.75	2.70	3.84	3.63
GBM-Stack	0.16	1.83	0.84	0.74	4.31	3.25	1.22	4.82	2.15
KNN	21.47	40.51	22.10	50.35	47.42	38.45	33.79	31.87	35.74
LR	0.82	6.33	9.27	23.38	25.95	7.38	10.97	23.46	13.44
LR-Stack	2.14	0.59	1.17	4.55	4.08	1.83	0.29	3.62	2.28
NB	53.71	37.68	6.02	102.40	70.14	62.94	7.89	55.28	49.51
NNET	1.66	0.11	1.79	7.78	7.05	11.04	6.81	13.78	6.25
RF	6.12	6.99	2.80	8.42	11.56	6.85	0.12	12.92	6.97
RLR	0.78	4.42	7.91	19.73	22.65	4.80	9.69	20.74	11.34
SC	23.72	25.85	11.90	29.09	23.68	25.59	16.16	19.39	21.92
Superlearner	2.98	1.19	3.50	2.51	3.76	3.49	1.47	1.63	2.57
SVM	10.83	0.83	3.60	1.59	5.78	10.46	6.52	0.06	4.96
Twang-GBM	13.25	14.85	6.08	17.58	13.98	16.37	5.06	12.17	12.42
Simulated PS	0.21	3.09	0.23	0.61	2.37	0.62	0.34	1.05	1.06

Table 48: (Mse; n=2000; ATE= -0.103) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0010	0.0010	0.0011	0.0012	0.0013	0.0012	0.0011	0.0011	0.0011
BAG-CART	0.0050	0.0103	0.0091	0.0079	0.0107	0.0095	0.0085	0.0090	0.0087
BOOSTLR	0.0030	0.0025	0.0037	0.0040	0.0035	0.0044	0.0031	0.0028	0.0034
GBM	0.0010	0.0009	0.0010	0.0010	0.0011	0.0009	0.0011	0.0010	0.0010
GBM-Stack	0.0010	0.0009	0.0010	0.0011	0.0012	0.0010	0.0011	0.0010	0.0010
KNN	0.0078	0.0093	0.0058	0.0104	0.0103	0.0110	0.0076	0.0061	0.0085
LR	0.0011	0.0009	0.0009	0.0013	0.0010	0.0010	0.0009	0.0009	0.0010
LR-Stack	0.0010	0.0009	0.0011	0.0010	0.0011	0.0011	0.0012	0.0010	0.0011
NB	0.0018	0.0015	0.0009	0.0022	0.0018	0.0016	0.0012	0.0015	0.0016
NNET	0.0011	0.0012	0.0014	0.0014	0.0016	0.0016	0.0014	0.0014	0.0014
RF	0.0013	0.0010	0.0011	0.0013	0.0013	0.0012	0.0011	0.0012	0.0012
RLR	0.0010	0.0009	0.0008	0.0012	0.0010	0.0010	0.0008	0.0009	0.0010
SC	0.0009	0.0008	0.0007	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008
Superlearner	0.0010	0.0009	0.0009	0.0010	0.0010	0.0009	0.0009	0.0009	0.0009
SVM	0.0012	0.0009	0.0011	0.0012	0.0012	0.0011	0.0011	0.0011	0.0011
Twang-GBM	0.0009	0.0008	0.0008	0.0010	0.0010	0.0009	0.0009	0.0009	0.0009
Simulated PS	0.0010	0.0012	0.0017	0.0014	0.0018	0.0014	0.0015	0.0017	0.0015

Table 49: (Mse; n=2000; ATE= -0.091) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0011	0.0011	0.0011	0.0013	0.0014	0.0013	0.0012	0.0012	0.0012
BAG-CART	0.0054	0.0109	0.0096	0.0084	0.0112	0.0101	0.0095	0.0095	0.0093
BOOSTLR	0.0032	0.0027	0.0040	0.0043	0.0039	0.0049	0.0035	0.0030	0.0037
GBM	0.0011	0.0010	0.0010	0.0011	0.0012	0.0010	0.0011	0.0011	0.0011
GBM-Stack	0.0010	0.0010	0.0010	0.0011	0.0013	0.0011	0.0012	0.0011	0.0011
KNN	0.0082	0.0099	0.0064	0.0108	0.0109	0.0119	0.0081	0.0067	0.0091
LR	0.0011	0.0010	0.0009	0.0014	0.0012	0.0011	0.0009	0.0010	0.0011
LR-Stack	0.0011	0.0010	0.0011	0.0011	0.0012	0.0012	0.0013	0.0011	0.0011
NB	0.0019	0.0016	0.0010	0.0026	0.0021	0.0018	0.0013	0.0017	0.0017
NNET	0.0011	0.0013	0.0016	0.0015	0.0018	0.0017	0.0016	0.0016	0.0015
RF	0.0013	0.0011	0.0012	0.0014	0.0014	0.0013	0.0013	0.0014	0.0013
RLR	0.0011	0.0009	0.0009	0.0013	0.0011	0.0011	0.0009	0.0010	0.0010
SC	0.0009	0.0009	0.0008	0.0010	0.0009	0.0009	0.0008	0.0009	0.0009
Superlearner	0.0010	0.0010	0.0009	0.0010	0.0011	0.0010	0.0010	0.0010	0.0010
SVM	0.0012	0.0010	0.0011	0.0013	0.0013	0.0012	0.0011	0.0012	0.0012
Twang-GBM	0.0010	0.0009	0.0009	0.0010	0.0011	0.0009	0.0009	0.0010	0.0010
Simulated PS	0.0011	0.0013	0.0017	0.0015	0.0019	0.0015	0.0016	0.0018	0.0016

Table 50: (Mse; n=2000; ATE= -0.078) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0012	0.0011	0.0013	0.0014	0.0015	0.0013	0.0013	0.0013	0.0013
BAG-CART	0.0063	0.0117	0.0102	0.0092	0.0121	0.0105	0.0102	0.0101	0.0100
BOOSTLR	0.0036	0.0029	0.0044	0.0047	0.0041	0.0052	0.0037	0.0034	0.0040
GBM	0.0012	0.0011	0.0012	0.0012	0.0013	0.0010	0.0012	0.0012	0.0012
GBM-Stack	0.0012	0.0011	0.0012	0.0012	0.0013	0.0011	0.0013	0.0012	0.0012
KNN	0.0088	0.0105	0.0067	0.0113	0.0117	0.0126	0.0091	0.0073	0.0097
LR	0.0013	0.0010	0.0011	0.0015	0.0013	0.0012	0.0010	0.0011	0.0012
LR-Stack	0.0012	0.0011	0.0013	0.0012	0.0013	0.0012	0.0014	0.0012	0.0012
NB	0.0022	0.0017	0.0011	0.0030	0.0022	0.0020	0.0014	0.0018	0.0019
NNET	0.0013	0.0014	0.0018	0.0016	0.0019	0.0018	0.0017	0.0018	0.0016
RF	0.0016	0.0012	0.0014	0.0015	0.0015	0.0015	0.0013	0.0015	0.0014
RLR	0.0013	0.0010	0.0011	0.0014	0.0013	0.0011	0.0010	0.0010	0.0011
SC	0.0010	0.0009	0.0009	0.0010	0.0010	0.0009	0.0009	0.0009	0.0010
Superlearner	0.0012	0.0010	0.0011	0.0011	0.0012	0.0010	0.0011	0.0011	0.0011
SVM	0.0013	0.0011	0.0013	0.0014	0.0014	0.0012	0.0012	0.0013	0.0013
Twang-GBM	0.0011	0.0010	0.0010	0.0011	0.0011	0.0010	0.0010	0.0010	0.0011
Simulated PS	0.0012	0.0014	0.0019	0.0016	0.0020	0.0016	0.0017	0.0019	0.0017

Table 51: (Mse; n=2000; ATE= -0.062) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0013	0.0012	0.0014	0.0016	0.0016	0.0015	0.0014	0.0015	0.0014
BAG-CART	0.0073	0.0128	0.0110	0.0099	0.0129	0.0120	0.0112	0.0110	0.0110
BOOSTLR	0.0039	0.0032	0.0047	0.0051	0.0045	0.0058	0.0040	0.0037	0.0044
GBM	0.0013	0.0012	0.0013	0.0013	0.0014	0.0011	0.0013	0.0014	0.0013
GBM-Stack	0.0012	0.0011	0.0013	0.0013	0.0014	0.0012	0.0014	0.0013	0.0013
KNN	0.0096	0.0113	0.0080	0.0119	0.0124	0.0131	0.0101	0.0079	0.0105
LR	0.0014	0.0011	0.0012	0.0017	0.0014	0.0013	0.0011	0.0012	0.0013
LR-Stack	0.0013	0.0011	0.0013	0.0013	0.0014	0.0014	0.0015	0.0014	0.0013
NB	0.0024	0.0018	0.0012	0.0036	0.0025	0.0023	0.0015	0.0022	0.0022
NNET	0.0014	0.0015	0.0019	0.0017	0.0021	0.0020	0.0019	0.0020	0.0018
RF	0.0017	0.0013	0.0015	0.0017	0.0017	0.0016	0.0015	0.0017	0.0016
RLR	0.0013	0.0011	0.0011	0.0016	0.0014	0.0012	0.0011	0.0012	0.0012
SC	0.0011	0.0010	0.0010	0.0011	0.0011	0.0010	0.0010	0.0011	0.0010
Superlearner	0.0013	0.0011	0.0011	0.0012	0.0013	0.0011	0.0011	0.0012	0.0012
SVM	0.0014	0.0011	0.0014	0.0015	0.0015	0.0013	0.0013	0.0014	0.0014
Twang-GBM	0.0012	0.0010	0.0011	0.0012	0.0012	0.0011	0.0011	0.0012	0.0011
Simulated PS	0.0013	0.0014	0.0020	0.0017	0.0021	0.0018	0.0018	0.0021	0.0018

Table 52: (Mse; n=2000; ATE= -0.044) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0014	0.0014	0.0015	0.0017	0.0018	0.0016	0.0016	0.0017	0.0016
BAG-CART	0.0082	0.0134	0.0125	0.0108	0.0144	0.0130	0.0126	0.0122	0.0121
BOOSTLR	0.0042	0.0035	0.0053	0.0053	0.0050	0.0064	0.0044	0.0042	0.0048
GBM	0.0014	0.0013	0.0014	0.0014	0.0015	0.0013	0.0015	0.0015	0.0014
GBM-Stack	0.0013	0.0013	0.0014	0.0014	0.0015	0.0014	0.0015	0.0015	0.0014
KNN	0.0102	0.0122	0.0089	0.0124	0.0135	0.0139	0.0107	0.0087	0.0113
LR	0.0015	0.0013	0.0013	0.0019	0.0017	0.0014	0.0012	0.0014	0.0014
LR-Stack	0.0014	0.0013	0.0015	0.0014	0.0015	0.0015	0.0016	0.0015	0.0015
NB	0.0026	0.0020	0.0013	0.0042	0.0029	0.0026	0.0016	0.0024	0.0024
NNET	0.0014	0.0016	0.0021	0.0019	0.0022	0.0022	0.0021	0.0023	0.0020
RF	0.0018	0.0015	0.0017	0.0019	0.0019	0.0018	0.0017	0.0019	0.0018
RLR	0.0014	0.0012	0.0013	0.0017	0.0016	0.0014	0.0012	0.0013	0.0014
SC	0.0012	0.0011	0.0010	0.0012	0.0012	0.0011	0.0011	0.0011	0.0011
Superlearner	0.0013	0.0012	0.0013	0.0013	0.0014	0.0013	0.0013	0.0014	0.0013
SVM	0.0015	0.0012	0.0015	0.0016	0.0016	0.0015	0.0015	0.0016	0.0015
Twang-GBM	0.0013	0.0012	0.0012	0.0013	0.0014	0.0012	0.0012	0.0013	0.0013
Simulated PS	0.0014	0.0016	0.0022	0.0019	0.0022	0.0020	0.0021	0.0022	0.0020

Table 53: (Mse; n=2000; ATE= 0.055) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0019	0.0018	0.0021	0.0025	0.0024	0.0024	0.0023	0.0025	0.0022
BAG-CART	0.0132	0.0178	0.0163	0.0166	0.0183	0.0192	0.0171	0.0172	0.0170
BOOSTLR	0.0058	0.0050	0.0076	0.0081	0.0072	0.0091	0.0062	0.0063	0.0069
GBM	0.0019	0.0017	0.0020	0.0020	0.0021	0.0018	0.0020	0.0021	0.0020
GBM-Stack	0.0019	0.0017	0.0020	0.0020	0.0021	0.0020	0.0021	0.0021	0.0020
KNN	0.0145	0.0157	0.0136	0.0160	0.0177	0.0178	0.0149	0.0127	0.0154
LR	0.0021	0.0018	0.0017	0.0030	0.0025	0.0021	0.0017	0.0022	0.0021
LR-Stack	0.0020	0.0016	0.0021	0.0021	0.0021	0.0021	0.0022	0.0022	0.0020
NB	0.0043	0.0028	0.0019	0.0081	0.0049	0.0043	0.0022	0.0043	0.0041
NNET	0.0021	0.0021	0.0030	0.0028	0.0030	0.0031	0.0032	0.0034	0.0028
RF	0.0029	0.0022	0.0025	0.0031	0.0030	0.0028	0.0026	0.0031	0.0028
RLR	0.0021	0.0017	0.0017	0.0028	0.0024	0.0020	0.0016	0.0021	0.0020
SC	0.0016	0.0015	0.0014	0.0017	0.0016	0.0016	0.0014	0.0016	0.0016
Superlearner	0.0019	0.0016	0.0018	0.0020	0.0020	0.0019	0.0018	0.0019	0.0019
SVM	0.0020	0.0016	0.0021	0.0021	0.0022	0.0021	0.0021	0.0022	0.0020
Twang-GBM	0.0018	0.0016	0.0017	0.0019	0.0019	0.0018	0.0017	0.0018	0.0018
Simulated PS	0.0020	0.0020	0.0031	0.0026	0.0027	0.0027	0.0029	0.0032	0.0027

Table 54: (MCse; n=2000; ATE= -0.103) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0320	0.0315	0.0328	0.0351	0.0364	0.0343	0.0333	0.0332	0.0336
BAG-CART	0.0701	0.0998	0.0953	0.0873	0.1019	0.0967	0.0921	0.0941	0.0922
BOOSTLR	0.0542	0.0500	0.0608	0.0633	0.0594	0.0655	0.0552	0.0530	0.0577
GBM	0.0320	0.0308	0.0317	0.0321	0.0331	0.0308	0.0328	0.0323	0.0319
GBM-Stack	0.0316	0.0307	0.0319	0.0326	0.0342	0.0318	0.0336	0.0318	0.0323
KNN	0.0885	0.0958	0.0754	0.1002	0.1010	0.1046	0.0872	0.0781	0.0913
LR	0.0329	0.0297	0.0293	0.0355	0.0322	0.0320	0.0294	0.0299	0.0314
LR-Stack	0.0322	0.0306	0.0329	0.0326	0.0338	0.0334	0.0347	0.0326	0.0329
NB	0.0392	0.0370	0.0306	0.0430	0.0404	0.0372	0.0351	0.0387	0.0377
NNET	0.0328	0.0342	0.0380	0.0369	0.0407	0.0401	0.0382	0.0382	0.0374
RF	0.0360	0.0318	0.0334	0.0357	0.0363	0.0352	0.0336	0.0352	0.0346
RLR	0.0325	0.0295	0.0291	0.0344	0.0317	0.0313	0.0292	0.0297	0.0309
SC	0.0286	0.0270	0.0271	0.0273	0.0276	0.0271	0.0276	0.0272	0.0274
Superlearner	0.0318	0.0297	0.0301	0.0311	0.0319	0.0304	0.0307	0.0301	0.0307
SVM	0.0343	0.0307	0.0328	0.0349	0.0352	0.0337	0.0331	0.0336	0.0335
Twang-GBM	0.0304	0.0285	0.0292	0.0304	0.0309	0.0291	0.0299	0.0294	0.0297
Simulated PS	0.0325	0.0346	0.0408	0.0376	0.0431	0.0383	0.0387	0.0409	0.0383

Table 55: (MCse; n=2000; ATE= -0.091) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0329	0.0330	0.0336	0.0366	0.0376	0.0354	0.0345	0.0348	0.0348
BAG-CART	0.0731	0.1026	0.0979	0.0899	0.1035	0.0997	0.0975	0.0961	0.0950
BOOSTLR	0.0558	0.0514	0.0632	0.0655	0.0622	0.0693	0.0583	0.0546	0.0600
GBM	0.0326	0.0321	0.0324	0.0333	0.0344	0.0316	0.0338	0.0336	0.0330
GBM-Stack	0.0324	0.0320	0.0325	0.0339	0.0353	0.0329	0.0344	0.0331	0.0333
KNN	0.0901	0.0991	0.0791	0.1020	0.1037	0.1090	0.0898	0.0817	0.0943
LR	0.0339	0.0309	0.0306	0.0369	0.0345	0.0333	0.0302	0.0314	0.0327
LR-Stack	0.0331	0.0318	0.0336	0.0338	0.0351	0.0344	0.0355	0.0338	0.0339
NB	0.0406	0.0382	0.0315	0.0451	0.0425	0.0390	0.0361	0.0399	0.0391
NNET	0.0338	0.0358	0.0396	0.0383	0.0420	0.0411	0.0398	0.0396	0.0388
RF	0.0368	0.0334	0.0352	0.0371	0.0378	0.0367	0.0353	0.0372	0.0362
RLR	0.0334	0.0307	0.0303	0.0358	0.0339	0.0326	0.0301	0.0311	0.0322
SC	0.0291	0.0281	0.0282	0.0283	0.0287	0.0279	0.0284	0.0281	0.0284
Superlearner	0.0325	0.0310	0.0309	0.0323	0.0333	0.0316	0.0315	0.0315	0.0318
SVM	0.0348	0.0319	0.0336	0.0359	0.0363	0.0347	0.0339	0.0350	0.0345
Twang-GBM	0.0311	0.0298	0.0299	0.0314	0.0322	0.0300	0.0307	0.0305	0.0307
Simulated PS	0.0335	0.0358	0.0418	0.0387	0.0440	0.0394	0.0401	0.0420	0.0394

Table 56: (MCse; n=2000; ATE= -0.078) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0348	0.0339	0.0359	0.0380	0.0392	0.0362	0.0357	0.0364	0.0363
BAG-CART	0.0783	0.1060	0.1009	0.0938	0.1076	0.1012	0.1012	0.0988	0.0985
BOOSTLR	0.0598	0.0533	0.0664	0.0685	0.0641	0.0714	0.0604	0.0576	0.0627
GBM	0.0347	0.0329	0.0346	0.0343	0.0355	0.0323	0.0348	0.0345	0.0342
GBM-Stack	0.0343	0.0329	0.0348	0.0350	0.0365	0.0336	0.0355	0.0343	0.0346
KNN	0.0935	0.1015	0.0814	0.1043	0.1070	0.1121	0.0947	0.0852	0.0975
LR	0.0359	0.0318	0.0327	0.0387	0.0359	0.0340	0.0316	0.0326	0.0342
LR-Stack	0.0350	0.0326	0.0355	0.0350	0.0362	0.0352	0.0367	0.0351	0.0352
NB	0.0427	0.0391	0.0336	0.0481	0.0439	0.0402	0.0368	0.0409	0.0406
NNET	0.0357	0.0367	0.0423	0.0397	0.0435	0.0424	0.0413	0.0417	0.0404
RF	0.0398	0.0347	0.0376	0.0387	0.0394	0.0385	0.0366	0.0382	0.0379
RLR	0.0354	0.0315	0.0324	0.0376	0.0353	0.0334	0.0314	0.0323	0.0337
SC	0.0305	0.0288	0.0298	0.0294	0.0296	0.0285	0.0296	0.0290	0.0294
Superlearner	0.0345	0.0319	0.0330	0.0334	0.0345	0.0324	0.0326	0.0326	0.0331
SVM	0.0365	0.0326	0.0356	0.0369	0.0373	0.0351	0.0351	0.0362	0.0357
Twang-GBM	0.0330	0.0307	0.0321	0.0324	0.0332	0.0307	0.0317	0.0316	0.0319
Simulated PS	0.0353	0.0367	0.0439	0.0399	0.0449	0.0401	0.0410	0.0434	0.0407

Table 57: (MCse; n=2000; ATE= -0.062) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0358	0.0350	0.0370	0.0401	0.0404	0.0382	0.0371	0.0388	0.0378
BAG-CART	0.0845	0.1111	0.1048	0.0970	0.1111	0.1081	0.1056	0.1033	0.1032
BOOSTLR	0.0623	0.0560	0.0687	0.0713	0.0670	0.0751	0.0625	0.0602	0.0654
GBM	0.0358	0.0341	0.0358	0.0360	0.0367	0.0335	0.0360	0.0368	0.0356
GBM-Stack	0.0353	0.0338	0.0359	0.0364	0.0375	0.0351	0.0368	0.0362	0.0359
KNN	0.0979	0.1057	0.0890	0.1070	0.1102	0.1138	0.0996	0.0887	0.1015
LR	0.0371	0.0333	0.0340	0.0409	0.0377	0.0357	0.0326	0.0350	0.0358
LR-Stack	0.0361	0.0335	0.0366	0.0366	0.0371	0.0367	0.0382	0.0370	0.0365
NB	0.0443	0.0405	0.0347	0.0518	0.0453	0.0425	0.0383	0.0446	0.0427
NNET	0.0369	0.0382	0.0435	0.0418	0.0453	0.0444	0.0432	0.0449	0.0423
RF	0.0413	0.0364	0.0390	0.0416	0.0411	0.0403	0.0385	0.0409	0.0399
RLR	0.0366	0.0329	0.0337	0.0397	0.0369	0.0349	0.0324	0.0345	0.0352
SC	0.0315	0.0297	0.0308	0.0306	0.0309	0.0296	0.0306	0.0309	0.0306
Superlearner	0.0354	0.0331	0.0340	0.0350	0.0356	0.0338	0.0339	0.0348	0.0345
SVM	0.0373	0.0335	0.0369	0.0384	0.0382	0.0364	0.0365	0.0380	0.0369
Twang-GBM	0.0340	0.0317	0.0331	0.0339	0.0345	0.0321	0.0329	0.0338	0.0332
Simulated PS	0.0366	0.0378	0.0452	0.0415	0.0458	0.0419	0.0425	0.0456	0.0421

Table 58: (MCse; n=2000; ATE= -0.044) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0371	0.0367	0.0390	0.0415	0.0429	0.0401	0.0397	0.0410	0.0398
BAG-CART	0.0894	0.1135	0.1117	0.1008	0.1174	0.1125	0.1121	0.1082	0.1082
BOOSTLR	0.0642	0.0583	0.0728	0.0723	0.0708	0.0789	0.0655	0.0643	0.0684
GBM	0.0372	0.0358	0.0376	0.0373	0.0387	0.0354	0.0383	0.0388	0.0374
GBM-Stack	0.0365	0.0355	0.0379	0.0376	0.0393	0.0370	0.0387	0.0384	0.0376
KNN	0.1006	0.1098	0.0936	0.1088	0.1148	0.1173	0.1026	0.0927	0.1050
LR	0.0384	0.0353	0.0355	0.0428	0.0409	0.0376	0.0344	0.0367	0.0377
LR-Stack	0.0374	0.0352	0.0385	0.0378	0.0390	0.0385	0.0402	0.0394	0.0383
NB	0.0459	0.0422	0.0365	0.0541	0.0484	0.0446	0.0403	0.0462	0.0448
NNET	0.0382	0.0398	0.0454	0.0440	0.0471	0.0467	0.0463	0.0474	0.0444
RF	0.0429	0.0384	0.0413	0.0438	0.0438	0.0427	0.0413	0.0437	0.0422
RLR	0.0379	0.0349	0.0352	0.0415	0.0400	0.0368	0.0342	0.0362	0.0371
SC	0.0326	0.0312	0.0321	0.0316	0.0327	0.0312	0.0321	0.0323	0.0320
Superlearner	0.0366	0.0348	0.0358	0.0364	0.0377	0.0357	0.0361	0.0368	0.0362
SVM	0.0385	0.0351	0.0388	0.0395	0.0402	0.0381	0.0388	0.0400	0.0386
Twang-GBM	0.0352	0.0333	0.0348	0.0350	0.0364	0.0338	0.0349	0.0355	0.0349
Simulated PS	0.0378	0.0394	0.0469	0.0431	0.0475	0.0445	0.0460	0.0475	0.0441

Table 59: (MCse; n=2000; ATE= 0.055) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0439	0.0423	0.0460	0.0501	0.0495	0.0491	0.0476	0.0498	0.0473
BAG-CART	0.1125	0.1302	0.1273	0.1248	0.1316	0.1359	0.1310	0.1275	0.1276
BOOSTLR	0.0751	0.0703	0.0873	0.0883	0.0846	0.0935	0.0776	0.0779	0.0818
GBM	0.0439	0.0416	0.0446	0.0449	0.0453	0.0429	0.0447	0.0458	0.0442
GBM-Stack	0.0431	0.0411	0.0446	0.0453	0.0462	0.0448	0.0456	0.0461	0.0446
KNN	0.1197	0.1233	0.1160	0.1233	0.1305	0.1317	0.1208	0.1113	0.1221
LR	0.0463	0.0419	0.0409	0.0536	0.0481	0.0459	0.0405	0.0448	0.0453
LR-Stack	0.0443	0.0408	0.0456	0.0459	0.0454	0.0461	0.0470	0.0464	0.0452
NB	0.0581	0.0485	0.0440	0.0693	0.0578	0.0554	0.0473	0.0575	0.0547
NNET	0.0458	0.0460	0.0550	0.0526	0.0543	0.0554	0.0563	0.0578	0.0529
RF	0.0540	0.0465	0.0497	0.0556	0.0546	0.0531	0.0510	0.0550	0.0524
RLR	0.0455	0.0414	0.0406	0.0520	0.0470	0.0448	0.0401	0.0441	0.0445
SC	0.0382	0.0364	0.0371	0.0382	0.0381	0.0373	0.0368	0.0382	0.0376
Superlearner	0.0435	0.0403	0.0423	0.0448	0.0445	0.0431	0.0426	0.0441	0.0432
SVM	0.0443	0.0403	0.0456	0.0463	0.0464	0.0450	0.0456	0.0464	0.0450
Twang-GBM	0.0413	0.0386	0.0410	0.0424	0.0429	0.0411	0.0407	0.0419	0.0412
Simulated PS	0.0451	0.0450	0.0555	0.0511	0.0525	0.0520	0.0543	0.0566	0.0515

Table 60: (SE; n=2000; ATE= -0.103) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0299	0.0317	0.0314	0.0331	0.0328	0.0331	0.0312	0.0320	0.0319
BAG-CART	0.0451	0.0637	0.0606	0.0557	0.0648	0.0611	0.0570	0.0604	0.0586
BOOSTLR	0.0447	0.0437	0.0469	0.0492	0.0465	0.0544	0.0435	0.0426	0.0464
GBM	0.0298	0.0308	0.0310	0.0311	0.0308	0.0305	0.0310	0.0316	0.0308
GBM-Stack	0.0297	0.0306	0.0313	0.0318	0.0317	0.0318	0.0317	0.0314	0.0312
KNN	0.0618	0.0663	0.0571	0.0741	0.0730	0.0759	0.0638	0.0580	0.0662
LR	0.0305	0.0290	0.0287	0.0315	0.0297	0.0311	0.0284	0.0287	0.0297
LR-Stack	0.0302	0.0306	0.0323	0.0319	0.0315	0.0329	0.0327	0.0316	0.0317
NB	0.0341	0.0339	0.0299	0.0358	0.0348	0.0353	0.0307	0.0330	0.0334
NNET	0.0302	0.0333	0.0348	0.0345	0.0353	0.0363	0.0344	0.0355	0.0343
RF	0.0315	0.0315	0.0314	0.0328	0.0322	0.0326	0.0311	0.0324	0.0319
RLR	0.0302	0.0288	0.0285	0.0311	0.0294	0.0307	0.0283	0.0286	0.0294
SC	0.0273	0.0268	0.0269	0.0274	0.0267	0.0270	0.0268	0.0270	0.0270
Superlearner	0.0297	0.0297	0.0295	0.0305	0.0299	0.0303	0.0294	0.0295	0.0298
SVM	0.0313	0.0306	0.0313	0.0334	0.0324	0.0331	0.0312	0.0317	0.0319
Twang-GBM	0.0289	0.0290	0.0290	0.0298	0.0292	0.0294	0.0289	0.0293	0.0292
Simulated PS	0.0302	0.0331	0.0349	0.0347	0.0358	0.0351	0.0341	0.0358	0.0342

Table 61: (SE; n=2000; ATE= -0.091) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0313	0.0328	0.0327	0.0349	0.0341	0.0345	0.0325	0.0335	0.0333
BAG-CART	0.0479	0.0659	0.0629	0.0587	0.0673	0.0639	0.0612	0.0631	0.0614
BOOSTLR	0.0466	0.0450	0.0490	0.0513	0.0484	0.0569	0.0456	0.0441	0.0484
GBM	0.0310	0.0318	0.0321	0.0325	0.0320	0.0317	0.0322	0.0329	0.0320
GBM-Stack	0.0310	0.0316	0.0324	0.0332	0.0328	0.0330	0.0328	0.0327	0.0324
KNN	0.0652	0.0697	0.0616	0.0771	0.0761	0.0793	0.0675	0.0611	0.0697
LR	0.0318	0.0300	0.0299	0.0332	0.0311	0.0323	0.0295	0.0300	0.0310
LR-Stack	0.0314	0.0316	0.0335	0.0333	0.0326	0.0341	0.0339	0.0329	0.0329
NB	0.0357	0.0350	0.0311	0.0377	0.0362	0.0366	0.0319	0.0345	0.0349
NNET	0.0316	0.0344	0.0364	0.0362	0.0366	0.0376	0.0359	0.0373	0.0358
RF	0.0329	0.0328	0.0330	0.0346	0.0336	0.0340	0.0327	0.0340	0.0334
RLR	0.0315	0.0298	0.0297	0.0327	0.0307	0.0319	0.0294	0.0298	0.0307
SC	0.0285	0.0277	0.0278	0.0285	0.0277	0.0280	0.0278	0.0281	0.0280
Superlearner	0.0310	0.0307	0.0306	0.0319	0.0310	0.0315	0.0305	0.0308	0.0310
SVM	0.0324	0.0316	0.0325	0.0347	0.0334	0.0342	0.0324	0.0330	0.0330
Twang-GBM	0.0302	0.0301	0.0302	0.0312	0.0303	0.0305	0.0300	0.0306	0.0304
Simulated PS	0.0315	0.0342	0.0364	0.0363	0.0370	0.0365	0.0355	0.0373	0.0356

Table 62: (SE; n=2000; ATE= -0.078) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0327	0.0339	0.0340	0.0366	0.0355	0.0360	0.0338	0.0351	0.0347
BAG-CART	0.0511	0.0687	0.0664	0.0621	0.0704	0.0670	0.0646	0.0659	0.0645
BOOSTLR	0.0484	0.0466	0.0514	0.0543	0.0505	0.0592	0.0478	0.0461	0.0505
GBM	0.0324	0.0329	0.0334	0.0341	0.0333	0.0330	0.0334	0.0343	0.0333
GBM-Stack	0.0323	0.0327	0.0337	0.0347	0.0341	0.0343	0.0341	0.0340	0.0337
KNN	0.0691	0.0730	0.0656	0.0808	0.0801	0.0824	0.0726	0.0651	0.0736
LR	0.0333	0.0311	0.0311	0.0350	0.0325	0.0338	0.0307	0.0315	0.0324
LR-Stack	0.0328	0.0327	0.0348	0.0347	0.0338	0.0354	0.0352	0.0342	0.0342
NB	0.0374	0.0362	0.0323	0.0402	0.0379	0.0382	0.0331	0.0363	0.0365
NNET	0.0331	0.0356	0.0381	0.0380	0.0380	0.0392	0.0374	0.0390	0.0373
RF	0.0346	0.0342	0.0345	0.0368	0.0354	0.0357	0.0343	0.0356	0.0351
RLR	0.0329	0.0309	0.0309	0.0345	0.0322	0.0333	0.0306	0.0313	0.0321
SC	0.0297	0.0287	0.0289	0.0298	0.0288	0.0291	0.0288	0.0293	0.0291
Superlearner	0.0324	0.0319	0.0319	0.0335	0.0324	0.0328	0.0318	0.0322	0.0323
SVM	0.0337	0.0327	0.0337	0.0362	0.0346	0.0354	0.0337	0.0343	0.0343
Twang-GBM	0.0315	0.0312	0.0314	0.0327	0.0316	0.0318	0.0312	0.0319	0.0317
Simulated PS	0.0329	0.0353	0.0379	0.0379	0.0384	0.0379	0.0369	0.0387	0.0370

Table 63: (SE; n=2000; ATE= -0.062) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0343	0.0354	0.0355	0.0385	0.0370	0.0377	0.0354	0.0370	0.0364
BAG-CART	0.0555	0.0723	0.0696	0.0654	0.0745	0.0712	0.0684	0.0701	0.0684
BOOSTLR	0.0510	0.0488	0.0538	0.0567	0.0531	0.0618	0.0498	0.0481	0.0529
GBM	0.0340	0.0343	0.0347	0.0357	0.0347	0.0345	0.0348	0.0359	0.0348
GBM-Stack	0.0339	0.0340	0.0351	0.0363	0.0355	0.0357	0.0356	0.0356	0.0352
KNN	0.0737	0.0777	0.0699	0.0851	0.0842	0.0862	0.0779	0.0699	0.0781
LR	0.0350	0.0327	0.0325	0.0370	0.0342	0.0354	0.0320	0.0331	0.0340
LR-Stack	0.0344	0.0341	0.0362	0.0364	0.0352	0.0368	0.0366	0.0358	0.0357
NB	0.0395	0.0377	0.0337	0.0430	0.0398	0.0400	0.0345	0.0386	0.0384
NNET	0.0347	0.0373	0.0397	0.0399	0.0397	0.0410	0.0393	0.0413	0.0391
RF	0.0366	0.0361	0.0362	0.0389	0.0373	0.0376	0.0360	0.0377	0.0371
RLR	0.0346	0.0324	0.0323	0.0365	0.0338	0.0349	0.0318	0.0328	0.0336
SC	0.0310	0.0299	0.0300	0.0312	0.0300	0.0304	0.0299	0.0307	0.0304
Superlearner	0.0340	0.0333	0.0332	0.0352	0.0338	0.0343	0.0331	0.0338	0.0338
SVM	0.0351	0.0339	0.0351	0.0377	0.0360	0.0368	0.0351	0.0358	0.0357
Twang-GBM	0.0330	0.0326	0.0327	0.0343	0.0331	0.0332	0.0325	0.0335	0.0331
Simulated PS	0.0346	0.0368	0.0395	0.0397	0.0398	0.0395	0.0385	0.0406	0.0386

Table 64: (SE; n=2000; ATE= -0.044) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0360	0.0369	0.0372	0.0404	0.0388	0.0395	0.0371	0.0390	0.0381
BAG-CART	0.0591	0.0757	0.0747	0.0694	0.0791	0.0753	0.0728	0.0735	0.0724
BOOSTLR	0.0534	0.0508	0.0566	0.0592	0.0560	0.0648	0.0523	0.0507	0.0555
GBM	0.0356	0.0357	0.0362	0.0373	0.0363	0.0360	0.0363	0.0376	0.0364
GBM-Stack	0.0355	0.0355	0.0367	0.0379	0.0370	0.0373	0.0371	0.0373	0.0368
KNN	0.0779	0.0819	0.0746	0.0889	0.0882	0.0901	0.0825	0.0746	0.0823
LR	0.0367	0.0342	0.0339	0.0390	0.0364	0.0371	0.0333	0.0348	0.0357
LR-Stack	0.0360	0.0355	0.0378	0.0380	0.0367	0.0384	0.0381	0.0375	0.0372
NB	0.0415	0.0394	0.0352	0.0455	0.0420	0.0420	0.0360	0.0408	0.0403
NNET	0.0364	0.0388	0.0417	0.0420	0.0414	0.0429	0.0413	0.0436	0.0410
RF	0.0389	0.0378	0.0382	0.0412	0.0394	0.0398	0.0379	0.0400	0.0391
RLR	0.0363	0.0340	0.0336	0.0384	0.0359	0.0365	0.0331	0.0345	0.0353
SC	0.0325	0.0312	0.0312	0.0326	0.0314	0.0317	0.0310	0.0321	0.0317
Superlearner	0.0356	0.0347	0.0346	0.0368	0.0354	0.0359	0.0346	0.0355	0.0354
SVM	0.0366	0.0353	0.0367	0.0392	0.0374	0.0383	0.0366	0.0374	0.0372
Twang-GBM	0.0346	0.0340	0.0341	0.0359	0.0346	0.0347	0.0339	0.0351	0.0346
Simulated PS	0.0363	0.0383	0.0411	0.0415	0.0414	0.0415	0.0404	0.0425	0.0404

Table 65: (SE; n=2000; ATE= 0.055) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=2000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0436	0.0432	0.0441	0.0499	0.0460	0.0478	0.0442	0.0478	0.0458
BAG-CART	0.0822	0.0930	0.0931	0.0940	0.1003	0.0955	0.0930	0.0942	0.0932
BOOSTLR	0.0647	0.0607	0.0677	0.0739	0.0683	0.0783	0.0645	0.0612	0.0674
GBM	0.0431	0.0418	0.0426	0.0452	0.0432	0.0428	0.0427	0.0452	0.0433
GBM-Stack	0.0428	0.0415	0.0432	0.0455	0.0436	0.0441	0.0435	0.0447	0.0436
KNN	0.1001	0.1017	0.0967	0.1091	0.1077	0.1098	0.1042	0.0952	0.1031
LR	0.0450	0.0412	0.0397	0.0503	0.0452	0.0453	0.0387	0.0432	0.0436
LR-Stack	0.0433	0.0415	0.0443	0.0457	0.0432	0.0452	0.0446	0.0449	0.0441
NB	0.0529	0.0467	0.0417	0.0613	0.0519	0.0521	0.0424	0.0515	0.0501
NNET	0.0445	0.0456	0.0498	0.0512	0.0493	0.0523	0.0504	0.0534	0.0496
RF	0.0495	0.0457	0.0463	0.0526	0.0492	0.0494	0.0462	0.0511	0.0487
RLR	0.0444	0.0408	0.0394	0.0492	0.0444	0.0444	0.0385	0.0427	0.0430
SC	0.0387	0.0365	0.0364	0.0389	0.0371	0.0373	0.0360	0.0384	0.0374
Superlearner	0.0432	0.0409	0.0408	0.0450	0.0424	0.0429	0.0407	0.0430	0.0424
SVM	0.0431	0.0410	0.0432	0.0463	0.0439	0.0447	0.0431	0.0444	0.0437
Twang-GBM	0.0417	0.0400	0.0402	0.0435	0.0413	0.0414	0.0399	0.0423	0.0413
Simulated PS	0.0444	0.0446	0.0485	0.0504	0.0486	0.0494	0.0486	0.0515	0.0482

Table 66: (Raw (naive) bias; n=2000) Summary table of relative ATE estimation bias (in %) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Relative bias is presented across all eight simulated scenarios (A-H) and across all six simulated average treatment effects based on the n=2000 data setup.

	A	B	C	D	E	F	G	H	Mean-bias
-1.2	17.99	20.00	3.28	24.43	22.87	19.76	6.54	19.78	16.83
-1.0	21.63	23.87	3.74	29.42	27.14	23.89	7.57	23.82	20.14
-0.8	26.70	29.70	4.71	38.10	33.95	29.99	9.41	30.07	25.33
-0.6	35.20	40.28	6.50	51.23	46.38	41.32	12.39	41.37	34.33
-0.4	53.98	61.45	10.11	77.84	71.15	63.62	18.80	62.84	52.47
+0.4	57.52	62.58	12.93	78.74	71.99	66.08	20.59	65.17	54.45

Table 67: (n=2000) Summary table of ATE estimation mean-squared error (mse) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. mean-squared error (mse) computed across m=1000 simulated data set based on the n=2000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-mse
-1.2	0.0011	0.0012	0.0007	0.0014	0.0013	0.0011	0.0008	0.0012	0.0011
-1.0	0.0012	0.0013	0.0008	0.0015	0.0014	0.0013	0.0008	0.0013	0.0012
-0.8	0.0013	0.0014	0.0009	0.0018	0.0016	0.0014	0.0009	0.0014	0.0013
-0.6	0.0015	0.0015	0.0009	0.0020	0.0018	0.0015	0.0010	0.0017	0.0015
-0.4	0.0016	0.0017	0.0010	0.0022	0.0021	0.0018	0.0011	0.0019	0.0017
+0.4	0.0025	0.0026	0.0014	0.0034	0.0031	0.0027	0.0015	0.0028	0.0025

Table 68: (n=2000) Summary table of ATE estimation Monte Carlo standard error (MCse) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Monte Carlo standard error (MCse) computed across m=1000 simulated data set based on the n=2000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-MCse
-1.2	0.0285	0.0272	0.0269	0.0274	0.0276	0.0271	0.0274	0.0278	0.0275
-1.0	0.0289	0.0284	0.0279	0.0285	0.0288	0.0280	0.0282	0.0288	0.0284
-0.8	0.0301	0.0290	0.0295	0.0296	0.0296	0.0285	0.0293	0.0298	0.0294
-0.6	0.0312	0.0299	0.0304	0.0306	0.0312	0.0297	0.0303	0.0316	0.0306
-0.4	0.0324	0.0314	0.0316	0.0316	0.0329	0.0313	0.0318	0.0329	0.0320
+0.4	0.0378	0.0364	0.0367	0.0380	0.0383	0.0369	0.0363	0.0385	0.0374

Table 69: (n=2000) Summary table of ATE estimation Robust sandwich-type standard error (SE) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Robust sandwich-type standard error (SE) computed across m=1000 simulated data set based on the n=2000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-bias
-1.2	0.0273	0.0268	0.0266	0.0276	0.0269	0.0270	0.0266	0.0274	0.0270
-1.0	0.0285	0.0278	0.0276	0.0288	0.0279	0.0281	0.0275	0.0286	0.0281
-0.8	0.0297	0.0289	0.0286	0.0302	0.0291	0.0292	0.0285	0.0300	0.0293
-0.6	0.0311	0.0301	0.0297	0.0315	0.0304	0.0305	0.0296	0.0314	0.0305
-0.4	0.0326	0.0314	0.0309	0.0330	0.0318	0.0319	0.0307	0.0329	0.0319
+0.4	0.0385	0.0365	0.0360	0.0388	0.0372	0.0373	0.0357	0.0388	0.0373

Table 70: (ASAM; n=2000) Mean of the averaged standardized absolute mean differences over m=1000 simulated data sets in the n=2000 simulated data setup. In each data set the mean of the standardized absolute mean differences of all ten covariates is taken. We describe values in this table as ASAM. The last row (NO WEIGHT) presents the ASAM in the initial non-weighted data.

	A	B	C	D	E	F	G	H	Mean-ASAM
AVNNET	0.106	0.103	0.103	0.129	0.112	0.119	0.103	0.117	0.111
BAG-CART	0.278	0.304	0.301	0.319	0.321	0.313	0.290	0.284	0.301
BOOSTLR	0.324	0.288	0.298	0.384	0.374	0.412	0.280	0.325	0.336
GBM	0.109	0.107	0.109	0.117	0.112	0.114	0.102	0.111	0.110
GBM-Stack	0.105	0.102	0.107	0.113	0.106	0.107	0.105	0.107	0.107
KNN	0.353	0.397	0.283	0.399	0.425	0.422	0.327	0.347	0.369
LR	0.107	0.113	0.096	0.158	0.144	0.118	0.092	0.129	0.120
LR-Stack	0.111	0.108	0.119	0.122	0.110	0.118	0.117	0.112	0.115
NB	0.211	0.167	0.111	0.364	0.240	0.230	0.114	0.194	0.204
NNET	0.105	0.111	0.121	0.133	0.123	0.139	0.124	0.140	0.125
RF	0.117	0.112	0.112	0.128	0.120	0.119	0.105	0.121	0.117
RLR	0.105	0.110	0.095	0.148	0.136	0.114	0.091	0.124	0.115
SC	0.191	0.163	0.120	0.206	0.186	0.212	0.126	0.151	0.169
Superlearner	0.105	0.103	0.100	0.114	0.107	0.107	0.096	0.103	0.104
SVM	0.117	0.105	0.105	0.124	0.110	0.114	0.106	0.111	0.112
Twang-GBM	0.108	0.103	0.092	0.115	0.111	0.115	0.087	0.105	0.104
Simulated PS	0.093	0.096	0.108	0.111	0.106	0.112	0.112	0.114	0.107
NO WEIGHT	0.291	0.264	0.154	0.342	0.313	0.326	0.159	0.257	0.263

Table 71: (ASAM$_{conf}$; n=2000) Mean of the averaged standardized absolute mean differences over m=1000 simulated data sets in the n=2000 simulated data setup. In each data set, the mean of the standardized absolute mean differences of the four confounding covariates is taken. We therefore describe values in this table with ASAM$_{conf}$. The last row (NO WEIGHT) presents the ASAM$_{conf}$ in the initial non-weighted data.

	A	B	C	D	E	F	G	H	Mean-ASAM
AVN	0.106	0.105	0.103	0.129	0.111	0.119	0.104	0.119	0.112
BAG-CART	0.276	0.317	0.296	0.330	0.333	0.337	0.279	0.295	0.308
BOOSTLR	0.323	0.278	0.339	0.478	0.472	0.534	0.323	0.382	0.391
GBM	0.113	0.111	0.113	0.125	0.117	0.126	0.108	0.112	0.116
GBM-Stack	0.104	0.102	0.107	0.115	0.107	0.112	0.108	0.108	0.108
KNN	0.338	0.398	0.287	0.398	0.450	0.443	0.335	0.352	0.375
LR	0.106	0.123	0.093	0.171	0.163	0.123	0.092	0.147	0.127
LR-Stack	0.111	0.106	0.118	0.126	0.110	0.123	0.121	0.114	0.116
NB	0.214	0.168	0.108	0.399	0.254	0.240	0.111	0.206	0.213
NNET	0.104	0.111	0.124	0.134	0.122	0.144	0.127	0.143	0.126
RF	0.114	0.108	0.118	0.124	0.115	0.119	0.105	0.119	0.115
RLR	0.105	0.120	0.092	0.159	0.152	0.119	0.091	0.140	0.122
SC	0.203	0.167	0.131	0.255	0.227	0.305	0.153	0.185	0.203
Superlearner	0.104	0.103	0.100	0.116	0.108	0.114	0.098	0.105	0.106
SVM	0.120	0.105	0.107	0.125	0.109	0.121	0.107	0.110	0.113
Twang-GBM	0.113	0.106	0.095	0.124	0.121	0.136	0.097	0.112	0.113
Simulated PS	0.093	0.098	0.112	0.115	0.108	0.114	0.113	0.116	0.109
NO WEIGHT	0.318	0.286	0.161	0.425	0.389	0.459	0.203	0.334	0.322

Table 72: Average of maximum IPTW weights resulting through the different propensity score estimation models in each of the m=1000 simulated data sets for each scenario (A-H) in the n=2000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVNNET	13.9	14.2	17.6	29.5	18.1	28.1	18.1	24.9	20.5
BAG-CART	180.7	182.1	184.1	192.5	188.6	188.8	181.9	189.2	186.0
BOOSTLR	66.5	54.4	96.1	109.4	91.4	124.7	78.2	62.1	85.3
GBM	15.1	13.0	15.2	15.7	14.5	13.9	14.8	16.2	14.8
GBM-Stack	10.5	9.7	11.2	11.5	10.9	11.5	12.0	11.4	11.1
KNN	199.9	199.8	198.2	199.9	200.0	200.0	199.7	200.0	199.7
LR	21.0	20.9	9.8	49.9	37.4	29.2	7.0	27.8	25.4
LR-Stack	12.9	9.9	15.1	14.1	10.6	15.7	15.1	11.7	13.1
NB	52.5	31.2	16.1	129.1	56.6	56.6	20.3	94.3	57.1
NNET	16.9	20.8	32.4	32.9	26.3	42.9	34.7	41.9	31.1
RF	42.7	29.5	29.3	47.3	39.2	40.8	26.9	45.7	37.7
RLR	19.3	19.3	9.3	44.2	33.5	26.0	6.6	25.1	22.9
SC	6.2	5.6	3.8	8.2	7.2	5.8	3.2	7.2	5.9
Superlearner	10.1	8.9	9.5	11.9	10.2	10.9	9.4	9.6	10.1
SVM	11.8	10.4	14.2	15.3	14.4	15.7	13.8	13.5	13.6
Twang-GBM	13.5	11.2	10.6	16.5	13.5	14.4	10.9	13.9	13.1
Simulated PS	19.3	19.9	31.6	27.6	25.6	30.3	33.8	31.4	27.4

Table 73: Average of mean IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=2000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVNNET	2.01	2.04	2.00	2.11	2.05	2.07	1.99	2.07	2.04
BAG-CART	3.27	3.27	3.33	3.64	3.49	3.45	3.23	3.39	3.38
BOOSTLR	3.20	3.06	3.22	3.74	3.56	3.79	3.08	3.16	3.35
GBM	2.00	2.00	1.96	1.98	1.99	1.96	1.97	2.01	1.98
GBM-Stack	2.00	2.01	1.99	2.01	2.02	2.01	2.01	2.01	2.01
KNN	6.33	6.02	5.63	6.66	6.55	6.25	5.76	6.58	6.22
LR	2.02	2.07	2.03	2.15	2.14	2.05	2.02	2.10	2.07
LR-Stack	2.03	2.04	2.07	2.05	2.02	2.09	2.06	2.02	2.05
NB	2.47	2.31	1.94	3.04	2.61	2.38	2.04	3.56	2.54
NNET	2.02	2.10	2.14	2.14	2.12	2.19	2.13	2.23	2.13
RF	2.27	2.19	2.15	2.23	2.20	2.17	2.10	2.25	2.19
RLR	2.00	2.05	2.02	2.11	2.11	2.02	2.01	2.08	2.05
SC	1.90	1.93	1.97	1.88	1.90	1.90	1.98	1.94	1.93
Superlearner	1.82	1.82	1.78	1.78	1.79	1.78	1.78	1.78	1.79
SVM	2.01	2.00	2.00	2.00	2.01	2.00	1.99	1.99	2.00
Twang-GBM	1.97	1.98	1.89	1.96	1.97	1.94	1.90	1.96	1.95
Simulated PS	2.00	2.00	2.00	2.00	2.00	2.00	1.99	1.99	2.00

Table 74: Average of the first quantile of the IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=2000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVN	1.20	1.24	1.23	1.14	1.18	1.17	1.23	1.17	1.19
BAG-CART	1.21	1.22	1.20	1.14	1.17	1.17	1.20	1.15	1.18
BOOSTLR	1.23	1.32	1.26	1.13	1.19	1.11	1.30	1.22	1.22
GBM	1.23	1.26	1.25	1.17	1.21	1.22	1.25	1.19	1.22
GBM-Stack	1.21	1.25	1.21	1.16	1.19	1.18	1.20	1.17	1.20
KNN	1.15	1.19	1.21	1.08	1.12	1.11	1.19	1.11	1.15
LR	1.21	1.31	1.45	1.17	1.23	1.21	1.48	1.25	1.29
LR-Stack	1.20	1.24	1.20	1.16	1.19	1.18	1.19	1.17	1.19
NB	1.13	1.18	1.27	1.10	1.13	1.15	1.27	1.06	1.16
NNET	1.20	1.22	1.20	1.14	1.17	1.16	1.19	1.15	1.18
RF	1.18	1.23	1.23	1.14	1.18	1.18	1.24	1.15	1.19
RLR	1.22	1.32	1.46	1.17	1.24	1.22	1.49	1.26	1.30
SC	1.34	1.51	1.58	1.33	1.42	1.43	1.63	1.36	1.45
SL	1.21	1.24	1.24	1.16	1.20	1.20	1.26	1.19	1.21
SVM	1.25	1.26	1.23	1.18	1.20	1.20	1.22	1.21	1.22
Twang-GBM	1.23	1.27	1.29	1.18	1.21	1.21	1.28	1.20	1.23
Simulated PS	1.21	1.23	1.20	1.13	1.17	1.15	1.18	1.13	1.18

Table 75: Average of the third quantile of the IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=2000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVN	2.11	2.18	2.12	1.94	2.03	2.02	2.10	1.99	2.06
BAG-CART	2.26	2.31	2.25	2.07	2.17	2.15	2.24	2.07	2.19
BOOSTLR	3.72	3.72	3.71	3.63	3.70	3.69	3.72	3.69	3.70
GBM	2.12	2.16	2.11	1.96	2.04	2.04	2.12	2.00	2.07
GBM-Stack	2.12	2.19	2.12	1.93	2.04	2.03	2.11	1.98	2.06
KNN	2.41	2.52	2.54	2.16	2.30	2.27	2.49	2.27	2.37
LR	2.10	2.17	2.24	1.97	2.05	2.08	2.29	2.09	2.12
LR-Stack	2.13	2.22	2.15	1.94	2.04	2.06	2.13	1.97	2.08
NB	2.18	2.24	2.11	2.03	2.10	2.16	2.21	2.32	2.17
NNET	2.10	2.18	2.15	1.95	2.04	2.05	2.13	2.02	2.08
RF	2.17	2.21	2.17	1.97	2.06	2.04	2.14	2.01	2.10
RLR	2.09	2.16	2.23	1.97	2.05	2.08	2.28	2.08	2.12
SC	2.27	2.13	2.39	2.13	2.10	2.15	2.36	2.32	2.23
SL	1.94	1.97	1.92	1.79	1.86	1.86	1.94	1.82	1.89
SVM	2.18	2.22	2.13	1.92	2.04	2.02	2.12	2.02	2.08
Twang-GBM	2.10	2.15	2.07	1.95	2.03	2.02	2.07	1.98	2.05
Simulated PS	2.08	2.12	2.05	1.88	1.97	1.94	2.02	1.86	1.99

12.3. Appendix A – Performance measure results based on n = 1000 setup

Table 76: (Relative Bias; n=1000; ATE= -0.103) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=1000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	0.88	1.60	1.87	0.74	2.23	0.22	1.52	2.26	1.42
BAG-CART	10.74	19.52	8.46	14.55	21.14	10.78	1.95	14.78	12.74
BOOSTLR	14.69	3.47	1.41	8.69	7.64	7.28	7.72	6.35	7.16
GBM	0.80	1.26	3.11	1.54	2.59	2.88	0.53	1.80	1.81
GBM-Stack	0.70	0.26	2.51	0.33	0.91	1.97	0.19	1.42	1.04
KNN	11.78	12.34	1.70	8.32	11.93	6.69	5.05	9.34	8.39
LR	0.77	1.70	1.35	1.32	0.24	0.83	2.64	2.99	1.48
LR-Stack	2.00	1.20	2.29	1.55	0.37	1.30	0.09	2.04	1.35
NB	14.87	10.38	3.25	20.13	13.26	14.83	1.17	7.51	10.67
NNET	0.20	1.18	0.33	0.18	0.47	1.45	3.66	6.89	1.79
RF	1.26	1.68	4.03	1.51	2.32	2.92	1.05	1.63	2.05
RLR	0.19	2.49	0.77	0.21	1.00	1.89	2.13	2.23	1.36
SC	6.81	10.46	4.14	10.63	10.90	10.22	4.44	6.38	8.00
Superlearner	0.36	3.44	2.49	2.60	3.83	3.94	0.69	1.33	2.33
SVM	0.35	1.67	0.96	0.21	1.51	3.16	1.72	0.88	1.31
Twang-GBM	4.21	7.00	3.80	8.16	8.48	8.01	1.85	4.60	5.76
Simulated PS	0.45	0.94	0.31	1.77	1.16	0.07	0.03	0.80	0.69

Table 77: (Relative Bias; n=1000; ATE= -0.091) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=1000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	1.24	1.56	1.98	0.59	3.75	0.25	2.68	3.07	1.89
BAG-CART	12.22	23.64	11.75	18.84	24.06	12.21	2.64	16.37	15.22
BOOSTLR	16.78	5.47	1.64	8.89	8.57	8.71	9.70	8.21	8.50
GBM	0.78	1.02	3.23	1.67	3.85	3.26	0.24	2.56	2.08
GBM-Stack	0.46	0.44	2.63	0.46	2.12	2.34	1.67	2.01	1.52
KNN	13.58	14.93	3.65	10.46	14.85	8.87	8.35	13.04	10.97
LR	0.89	1.25	1.67	2.29	0.17	0.89	4.06	4.45	1.96
LR-Stack	1.94	1.02	2.17	1.64	1.28	1.57	1.20	2.69	1.69
NB	18.19	13.12	3.56	25.49	15.49	17.72	0.66	10.91	13.14
NNET	0.48	1.99	0.52	0.59	1.33	1.96	5.25	8.70	2.60
RF	1.24	1.45	4.62	1.81	3.46	3.60	0.22	1.96	2.30
RLR	0.14	2.24	1.00	0.77	1.40	2.14	3.43	3.45	1.82
SC	8.56	12.22	4.76	12.90	13.47	12.08	4.61	7.54	9.52
Superlearner	0.78	3.58	2.69	3.10	5.23	4.63	0.25	1.19	2.68
SVM	1.22	1.73	0.77	0.58	2.72	3.81	3.25	1.19	1.91
Twang-GBM	5.35	7.88	4.22	9.81	10.86	9.41	1.22	5.27	6.75
Simulated PS	0.48	0.78	0.45	2.17	2.19	0.12	0.82	1.45	1.06

Table 78: (Relative Bias; n=1000; ATE= -0.078) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=1000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	1.47	1.84	1.61	0.96	4.12	0.99	4.41	3.44	2.36
BAG-CART	16.07	29.23	14.07	24.97	26.05	13.43	5.21	18.63	18.46
BOOSTLR	19.56	6.67	2.50	12.03	10.73	10.94	12.45	11.08	10.75
GBM	0.88	1.25	3.55	1.44	4.33	3.58	0.87	3.08	2.37
GBM-Stack	0.29	0.68	2.92	0.25	2.31	2.46	2.84	1.78	1.69
KNN	15.26	18.83	2.97	13.31	15.95	13.19	10.20	18.49	13.52
LR	1.36	0.66	2.37	4.24	0.99	0.37	6.28	6.05	2.79
LR-Stack	2.09	0.96	2.38	2.63	1.27	1.65	2.45	2.63	2.01
NB	23.73	16.33	4.37	34.10	20.12	23.45	0.20	14.69	17.12
NNET	0.45	2.62	1.41	1.01	0.96	3.17	7.11	10.04	3.35
RF	1.86	1.43	5.13	1.32	4.03	4.18	0.08	2.35	2.55
RLR	0.33	1.97	1.49	2.19	1.01	1.98	5.46	4.63	2.38
SC	11.08	14.98	5.98	15.69	16.25	14.52	4.97	9.83	11.66
Superlearner	1.06	4.19	2.90	3.22	6.15	5.26	1.16	1.69	3.20
SVM	1.99	2.05	0.45	0.82	3.04	4.40	4.78	0.57	2.26
Twang-GBM	6.85	9.91	5.04	11.93	13.31	11.47	0.93	6.85	8.29
Simulated PS	0.63	1.11	1.23	2.31	2.80	0.60	1.68	1.33	1.46

Table 79: (Relative Bias; n=1000; ATE= -0.062) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=1000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	2.30	2.29	1.68	3.01	5.04	1.31	7.36	4.95	3.49
BAG-CART	24.03	38.43	19.69	35.64	32.48	18.62	5.04	22.52	24.56
BOOSTLR	25.46	8.14	5.55	17.09	14.38	16.73	17.11	13.59	14.76
GBM	0.94	1.42	3.93	0.81	4.98	4.83	2.70	3.89	2.94
GBM-Stack	0.30	0.93	3.06	0.22	2.45	3.26	5.19	1.66	2.13
KNN	22.45	26.34	5.68	20.20	21.04	16.09	14.32	24.33	18.81
LR	1.26	0.07	4.42	8.40	3.53	0.17	9.95	9.18	4.62
LR-Stack	1.85	1.16	2.44	4.15	1.00	3.25	4.57	2.88	2.66
NB	31.27	21.98	5.57	49.42	29.53	32.39	0.86	22.03	24.13
NNET	1.08	4.22	2.42	3.08	0.32	4.19	11.84	12.09	4.90
RF	2.73	0.73	5.56	0.05	4.95	4.75	1.18	2.46	2.80
RLR	0.13	1.77	3.17	5.27	0.68	2.00	8.84	7.10	3.62
SC	14.92	19.75	7.33	20.40	21.28	18.85	5.63	13.43	15.20
Superlearner	1.60	5.18	3.18	3.09	7.29	6.77	2.84	2.62	4.07
SVM	3.29	2.53	0.10	0.46	3.55	6.29	7.45	0.32	3.00
Twang-GBM	9.27	12.70	6.13	15.23	17.32	15.26	0.16	9.61	10.71
Simulated PS	0.27	1.47	2.04	2.62	3.17	0.22	4.06	0.29	1.77

Table 80: (Relative Bias; n=1000; ATE= -0.044) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=1000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	3.09	1.99	1.60	3.75	7.50	2.88	12.12	8.46	5.18
BAG-CART	37.75	56.55	29.04	52.22	48.02	30.79	6.17	31.98	36.56
BOOSTLR	37.48	12.37	7.99	27.66	20.27	26.31	25.58	22.18	22.48
GBM	1.47	2.03	5.16	2.25	7.68	5.98	4.24	4.79	4.20
GBM-Stack	0.12	1.98	3.24	0.49	3.95	3.61	8.16	1.46	2.88
KNN	31.43	43.16	11.48	27.26	35.37	29.09	17.86	39.56	29.40
LR	2.80	1.92	7.85	12.91	8.73	2.27	15.69	15.50	8.46
LR-Stack	2.77	0.64	1.97	5.55	1.65	3.78	7.55	3.27	3.40
NB	50.16	33.16	7.21	74.78	47.62	51.58	1.22	36.85	37.82
NNET	1.09	6.62	4.86	2.17	0.41	6.76	17.85	19.64	7.42
RF	4.63	0.36	7.01	0.85	7.04	5.31	2.36	4.60	4.02
RLR	0.46	1.02	5.92	8.04	4.15	1.22	13.89	12.07	5.85
SC	23.05	28.78	10.56	31.71	31.73	27.70	8.28	21.07	22.86
Superlearner	2.29	6.99	3.82	5.48	10.71	8.96	4.52	4.14	5.86
SVM	6.04	2.90	0.70	1.55	6.06	8.64	11.23	0.85	4.75
Twang-GBM	14.25	18.79	8.34	23.65	26.02	22.37	1.22	16.43	16.39
Simulated PS	1.13	2.25	4.02	4.87	5.23	2.11	7.20	0.14	3.37

Table 81: (Relative Bias; n=1000; ATE= 0.055) Absolute relative bias of the ATE estimate (in %) computed over m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=1000 simulated data setup. The average bias across all scenarios (A-H) is presented in the last column. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting bias by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean bias
AVNNET	3.26	1.00	0.24	6.37	6.09	4.39	10.72	10.59	5.33
BAG-CART	45.40	51.69	20.34	61.82	47.55	29.26	6.43	33.03	36.94
BOOSTLR	32.86	10.66	4.09	39.56	28.48	28.07	23.51	19.73	23.37
GBM	1.27	0.03	6.07	2.16	5.64	8.79	4.66	5.50	4.27
GBM-Stack	1.11	4.90	2.71	2.98	2.20	7.05	7.43	2.00	3.80
KNN	32.44	47.66	17.99	41.22	37.45	31.89	26.02	45.90	35.07
LR	2.76	11.12	8.72	24.51	19.52	4.07	15.81	22.14	13.58
LR-Stack	1.93	2.64	0.54	9.51	1.11	5.93	7.69	4.30	4.21
NB	56.98	39.03	6.25	96.31	59.43	56.69	0.62	50.81	45.76
NNET	0.63	6.81	9.39	7.67	1.53	5.52	16.72	19.76	8.50
RF	5.56	1.54	5.64	5.13	5.72	7.60	4.35	6.70	5.28
RLR	0.21	7.65	6.34	18.36	13.73	0.10	13.66	18.01	9.76
SC	24.66	24.09	12.66	28.55	29.09	28.94	10.76	19.92	22.33
Superlearner	2.94	3.24	4.04	0.21	7.74	10.02	4.59	3.02	4.47
SVM	10.25	0.40	2.12	0.13	4.05	12.75	11.18	1.35	5.28
Twang-GBM	15.21	16.75	7.83	20.88	25.37	24.68	1.03	17.44	16.15
Simulated PS	1.24	0.92	5.43	2.21	5.80	0.81	6.47	1.36	3.03

Table 82: (Mse; n=1000; ATE= -0.103) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0018	0.0024	0.0028	0.0026	0.0026	0.0029	0.0020	0.0026	0.0024
BAG-CART	0.0099	0.0157	0.0162	0.0097	0.0163	0.0108	0.0146	0.0115	0.0131
BOOSTLR	0.0069	0.0056	0.0075	0.0082	0.0089	0.0086	0.0066	0.0063	0.0073
GBM	0.0019	0.0019	0.0021	0.0021	0.0021	0.0019	0.0020	0.0020	0.0020
GBM-Stack	0.0018	0.0018	0.0021	0.0021	0.0021	0.0020	0.0020	0.0019	0.0020
KNN	0.0127	0.0154	0.0108	0.0147	0.0149	0.0196	0.0111	0.0112	0.0138
LR	0.0020	0.0016	0.0017	0.0024	0.0018	0.0021	0.0016	0.0018	0.0019
LR-Stack	0.0019	0.0018	0.0023	0.0024	0.0023	0.0023	0.0023	0.0020	0.0022
NB	0.0032	0.0027	0.0020	0.0038	0.0029	0.0030	0.0020	0.0027	0.0028
NNET	0.0018	0.0027	0.0037	0.0028	0.0028	0.0031	0.0030	0.0037	0.0029
RF	0.0020	0.0019	0.0023	0.0024	0.0022	0.0021	0.0020	0.0022	0.0021
RLR	0.0019	0.0016	0.0016	0.0023	0.0017	0.0019	0.0016	0.0017	0.0018
SC	0.0015	0.0014	0.0015	0.0016	0.0015	0.0015	0.0014	0.0015	0.0015
Superlearner	0.0017	0.0016	0.0017	0.0019	0.0018	0.0018	0.0016	0.0017	0.0017
SVM	0.0021	0.0017	0.0018	0.0023	0.0021	0.0022	0.0018	0.0019	0.0020
Twang-GBM	0.0016	0.0015	0.0017	0.0018	0.0017	0.0017	0.0016	0.0017	0.0017
Simulated PS	0.0018	0.0024	0.0026	0.0027	0.0027	0.0031	0.0025	0.0034	0.0027

Table 83: (Mse; n=1000; ATE= -0.091) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0019	0.0025	0.0029	0.0028	0.0028	0.0030	0.0022	0.0027	0.0026
BAG-CART	0.0113	0.0160	0.0166	0.0103	0.0171	0.0121	0.0154	0.0130	0.0140
BOOSTLR	0.0074	0.0058	0.0082	0.0092	0.0095	0.0092	0.0071	0.0068	0.0079
GBM	0.0021	0.0021	0.0023	0.0023	0.0022	0.0021	0.0021	0.0022	0.0022
GBM-Stack	0.0019	0.0019	0.0023	0.0022	0.0022	0.0021	0.0021	0.0021	0.0021
KNN	0.0140	0.0166	0.0118	0.0155	0.0160	0.0200	0.0127	0.0118	0.0148
LR	0.0022	0.0017	0.0018	0.0027	0.0019	0.0023	0.0017	0.0020	0.0020
LR-Stack	0.0020	0.0019	0.0025	0.0025	0.0025	0.0024	0.0024	0.0022	0.0023
NB	0.0035	0.0028	0.0022	0.0042	0.0032	0.0034	0.0022	0.0030	0.0030
NNET	0.0020	0.0028	0.0040	0.0030	0.0029	0.0032	0.0032	0.0040	0.0031
RF	0.0022	0.0020	0.0025	0.0025	0.0023	0.0023	0.0022	0.0025	0.0023
RLR	0.0021	0.0017	0.0018	0.0025	0.0019	0.0021	0.0016	0.0019	0.0019
SC	0.0016	0.0015	0.0016	0.0018	0.0016	0.0016	0.0015	0.0016	0.0016
Superlearner	0.0019	0.0017	0.0018	0.0020	0.0019	0.0019	0.0017	0.0019	0.0018
SVM	0.0022	0.0019	0.0020	0.0025	0.0022	0.0023	0.0019	0.0021	0.0021
Twang-GBM	0.0018	0.0016	0.0018	0.0019	0.0018	0.0018	0.0017	0.0018	0.0018
Simulated PS	0.0020	0.0025	0.0028	0.0028	0.0028	0.0033	0.0027	0.0037	0.0028

Table 84: (Mse; n=1000; ATE= -0.078) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0022	0.0026	0.0030	0.0032	0.0030	0.0034	0.0024	0.0030	0.0028
BAG-CART	0.0126	0.0170	0.0176	0.0118	0.0188	0.0138	0.0165	0.0143	0.0153
BOOSTLR	0.0083	0.0062	0.0091	0.0101	0.0102	0.0104	0.0078	0.0074	0.0087
GBM	0.0023	0.0021	0.0025	0.0026	0.0023	0.0023	0.0022	0.0024	0.0023
GBM-Stack	0.0021	0.0019	0.0025	0.0024	0.0023	0.0024	0.0023	0.0022	0.0023
KNN	0.0152	0.0177	0.0127	0.0172	0.0168	0.0207	0.0136	0.0128	0.0158
LR	0.0024	0.0019	0.0020	0.0032	0.0022	0.0025	0.0018	0.0022	0.0023
LR-Stack	0.0022	0.0019	0.0027	0.0028	0.0026	0.0026	0.0026	0.0024	0.0025
NB	0.0039	0.0030	0.0024	0.0052	0.0035	0.0037	0.0023	0.0033	0.0034
NNET	0.0022	0.0030	0.0042	0.0033	0.0031	0.0035	0.0034	0.0043	0.0034
RF	0.0025	0.0021	0.0028	0.0028	0.0025	0.0026	0.0023	0.0027	0.0025
RLR	0.0023	0.0018	0.0019	0.0029	0.0021	0.0024	0.0018	0.0021	0.0021
SC	0.0018	0.0016	0.0017	0.0019	0.0017	0.0018	0.0015	0.0018	0.0017
Superlearner	0.0021	0.0017	0.0020	0.0022	0.0020	0.0022	0.0018	0.0020	0.0020
SVM	0.0024	0.0019	0.0022	0.0027	0.0023	0.0025	0.0020	0.0022	0.0023
Twang-GBM	0.0020	0.0017	0.0020	0.0022	0.0019	0.0020	0.0018	0.0020	0.0020
Simulated PS	0.0022	0.0026	0.0031	0.0031	0.0030	0.0036	0.0029	0.0038	0.0030

Table 85: (Mse; n=1000; ATE= -0.062) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0023	0.0028	0.0033	0.0034	0.0033	0.0038	0.0027	0.0034	0.0031
BAG-CART	0.0136	0.0180	0.0189	0.0132	0.0208	0.0154	0.0176	0.0163	0.0167
BOOSTLR	0.0088	0.0066	0.0104	0.0111	0.0110	0.0113	0.0085	0.0083	0.0095
GBM	0.0024	0.0022	0.0028	0.0028	0.0026	0.0026	0.0025	0.0027	0.0026
GBM-Stack	0.0023	0.0021	0.0028	0.0026	0.0025	0.0026	0.0026	0.0026	0.0025
KNN	0.0162	0.0190	0.0146	0.0183	0.0184	0.0222	0.0153	0.0144	0.0173
LR	0.0026	0.0020	0.0022	0.0034	0.0025	0.0028	0.0020	0.0025	0.0025
LR-Stack	0.0024	0.0021	0.0030	0.0029	0.0028	0.0030	0.0028	0.0027	0.0027
NB	0.0043	0.0032	0.0027	0.0058	0.0040	0.0042	0.0026	0.0038	0.0038
NNET	0.0023	0.0032	0.0047	0.0035	0.0033	0.0038	0.0041	0.0048	0.0037
RF	0.0027	0.0023	0.0033	0.0030	0.0029	0.0029	0.0026	0.0031	0.0029
RLR	0.0025	0.0019	0.0022	0.0031	0.0023	0.0026	0.0020	0.0024	0.0024
SC	0.0019	0.0018	0.0019	0.0021	0.0019	0.0020	0.0017	0.0020	0.0019
Superlearner	0.0022	0.0019	0.0022	0.0024	0.0022	0.0024	0.0020	0.0023	0.0022
SVM	0.0025	0.0021	0.0025	0.0029	0.0025	0.0028	0.0023	0.0026	0.0025
Twang-GBM	0.0022	0.0019	0.0022	0.0024	0.0021	0.0023	0.0021	0.0023	0.0022
Simulated PS	0.0024	0.0028	0.0034	0.0034	0.0033	0.0039	0.0033	0.0043	0.0033

Table 86: (Mse; n=1000; ATE= -0.044) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0025	0.0030	0.0037	0.0038	0.0034	0.0042	0.0030	0.0037	0.0034
BAG-CART	0.0154	0.0190	0.0207	0.0150	0.0220	0.0167	0.0194	0.0178	0.0182
BOOSTLR	0.0096	0.0072	0.0112	0.0120	0.0117	0.0123	0.0091	0.0089	0.0103
GBM	0.0027	0.0024	0.0031	0.0031	0.0027	0.0027	0.0026	0.0031	0.0028
GBM-Stack	0.0025	0.0022	0.0031	0.0029	0.0026	0.0028	0.0027	0.0028	0.0027
KNN	0.0177	0.0203	0.0162	0.0204	0.0195	0.0237	0.0164	0.0160	0.0188
LR	0.0028	0.0022	0.0025	0.0039	0.0026	0.0030	0.0022	0.0028	0.0028
LR-Stack	0.0026	0.0022	0.0033	0.0032	0.0030	0.0031	0.0030	0.0030	0.0029
NB	0.0047	0.0036	0.0029	0.0068	0.0042	0.0047	0.0027	0.0042	0.0042
NNET	0.0025	0.0036	0.0052	0.0038	0.0035	0.0040	0.0045	0.0052	0.0040
RF	0.0030	0.0025	0.0036	0.0035	0.0031	0.0031	0.0029	0.0034	0.0031
RLR	0.0027	0.0021	0.0025	0.0035	0.0024	0.0028	0.0021	0.0027	0.0026
SC	0.0021	0.0019	0.0021	0.0023	0.0020	0.0021	0.0018	0.0022	0.0021
Superlearner	0.0024	0.0021	0.0025	0.0027	0.0023	0.0026	0.0022	0.0026	0.0024
SVM	0.0028	0.0022	0.0027	0.0032	0.0027	0.0029	0.0024	0.0028	0.0027
Twang-GBM	0.0024	0.0020	0.0024	0.0027	0.0022	0.0024	0.0022	0.0026	0.0024
Simulated PS	0.0026	0.0029	0.0037	0.0037	0.0033	0.0040	0.0035	0.0046	0.0036

Table 87: (Mse; n=1000; ATE= 0.055) Mean squared error (mse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting mse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean mse
AVNNET	0.0036	0.0039	0.0052	0.0064	0.0049	0.0061	0.0042	0.0055	0.0050
BAG-CART	0.0233	0.0255	0.0286	0.0263	0.0293	0.0260	0.0263	0.0256	0.0264
BOOSTLR	0.0142	0.0111	0.0149	0.0167	0.0156	0.0181	0.0120	0.0135	0.0145
GBM	0.0038	0.0035	0.0040	0.0047	0.0038	0.0040	0.0035	0.0041	0.0039
GBM-Stack	0.0035	0.0032	0.0041	0.0045	0.0035	0.0041	0.0037	0.0038	0.0038
KNN	0.0251	0.0274	0.0245	0.0280	0.0269	0.0307	0.0235	0.0233	0.0262
LR	0.0042	0.0034	0.0034	0.0068	0.0043	0.0049	0.0029	0.0044	0.0043
LR-Stack	0.0037	0.0032	0.0043	0.0051	0.0040	0.0046	0.0040	0.0040	0.0041
NB	0.0073	0.0051	0.0039	0.0133	0.0070	0.0078	0.0039	0.0068	0.0069
NNET	0.0037	0.0052	0.0072	0.0059	0.0054	0.0061	0.0067	0.0078	0.0060
RF	0.0046	0.0040	0.0049	0.0059	0.0045	0.0049	0.0042	0.0053	0.0048
RLR	0.0039	0.0033	0.0032	0.0060	0.0038	0.0045	0.0029	0.0041	0.0040
SC	0.0030	0.0028	0.0027	0.0034	0.0028	0.0032	0.0024	0.0031	0.0029
Superlearner	0.0034	0.0030	0.0033	0.0043	0.0032	0.0039	0.0030	0.0036	0.0035
SVM	0.0038	0.0032	0.0037	0.0047	0.0036	0.0043	0.0034	0.0039	0.0038
Twang-GBM	0.0034	0.0030	0.0032	0.0041	0.0032	0.0037	0.0030	0.0035	0.0034
Simulated PS	0.0038	0.0040	0.0050	0.0056	0.0044	0.0055	0.0049	0.0064	0.0049

Table 88: (MCse; n=1000; ATE= -0.103) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0422	0.0487	0.0528	0.0510	0.0512	0.0536	0.0452	0.0506	0.0494
BAG-CART	0.0990	0.1237	0.1269	0.0975	0.1263	0.1032	0.1208	0.1063	0.1130
BOOSTLR	0.0822	0.0750	0.0868	0.0904	0.0940	0.0925	0.0809	0.0795	0.0852
GBM	0.0435	0.0441	0.0460	0.0463	0.0461	0.0440	0.0449	0.0453	0.0450
GBM-Stack	0.0423	0.0423	0.0457	0.0455	0.0457	0.0449	0.0448	0.0439	0.0444
KNN	0.1121	0.1237	0.1041	0.1212	0.1220	0.1398	0.1052	0.1054	0.1167
LR	0.0450	0.0406	0.0407	0.0497	0.0426	0.0454	0.0403	0.0424	0.0433
LR-Stack	0.0431	0.0422	0.0482	0.0489	0.0485	0.0476	0.0477	0.0452	0.0464
NB	0.0545	0.0507	0.0448	0.0582	0.0521	0.0529	0.0454	0.0516	0.0513
NNET	0.0427	0.0522	0.0609	0.0534	0.0527	0.0555	0.0548	0.0604	0.0541
RF	0.0444	0.0436	0.0480	0.0487	0.0468	0.0454	0.0448	0.0473	0.0461
RLR	0.0438	0.0398	0.0403	0.0478	0.0417	0.0439	0.0398	0.0416	0.0423
SC	0.0381	0.0364	0.0380	0.0391	0.0376	0.0376	0.0377	0.0384	0.0379
Superlearner	0.0415	0.0396	0.0408	0.0433	0.0419	0.0426	0.0401	0.0413	0.0414
SVM	0.0455	0.0417	0.0429	0.0485	0.0459	0.0467	0.0421	0.0442	0.0447
Twang-GBM	0.0403	0.0389	0.0407	0.0423	0.0406	0.0406	0.0404	0.0406	0.0406
Simulated PS	0.0425	0.0489	0.0511	0.0516	0.0525	0.0561	0.0501	0.0585	0.0514

Table 89: (MCse; n=1000; ATE= -0.091) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0440	0.0499	0.0542	0.0535	0.0531	0.0551	0.0467	0.0522	0.0511
BAG-CART	0.1058	0.1249	0.1284	0.1002	0.1293	0.1095	0.1241	0.1130	0.1169
BOOSTLR	0.0849	0.0762	0.0907	0.0957	0.0975	0.0956	0.0841	0.0822	0.0884
GBM	0.0454	0.0454	0.0482	0.0478	0.0472	0.0455	0.0461	0.0472	0.0466
GBM-Stack	0.0440	0.0435	0.0478	0.0469	0.0473	0.0461	0.0461	0.0456	0.0459
KNN	0.1175	0.1284	0.1086	0.1243	0.1259	0.1414	0.1125	0.1082	0.1208
LR	0.0471	0.0419	0.0423	0.0520	0.0443	0.0476	0.0409	0.0445	0.0451
LR-Stack	0.0447	0.0433	0.0502	0.0502	0.0501	0.0489	0.0492	0.0467	0.0479
NB	0.0568	0.0519	0.0468	0.0608	0.0545	0.0558	0.0466	0.0537	0.0534
NNET	0.0446	0.0534	0.0635	0.0550	0.0543	0.0568	0.0567	0.0626	0.0559
RF	0.0467	0.0450	0.0501	0.0505	0.0484	0.0476	0.0465	0.0505	0.0482
RLR	0.0457	0.0411	0.0419	0.0500	0.0433	0.0458	0.0404	0.0436	0.0440
SC	0.0399	0.0377	0.0396	0.0404	0.0387	0.0387	0.0382	0.0398	0.0391
Superlearner	0.0434	0.0409	0.0425	0.0447	0.0432	0.0439	0.0412	0.0431	0.0429
SVM	0.0472	0.0431	0.0449	0.0498	0.0474	0.0478	0.0434	0.0460	0.0462
Twang-GBM	0.0423	0.0402	0.0424	0.0434	0.0416	0.0419	0.0414	0.0424	0.0420
Simulated PS	0.0446	0.0501	0.0532	0.0531	0.0534	0.0578	0.0520	0.0604	0.0531

Table 90: (MCse; n=1000; ATE= -0.078) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0464	0.0514	0.0551	0.0565	0.0548	0.0582	0.0486	0.0544	0.0532
BAG-CART	0.1114	0.1284	0.1322	0.1069	0.1359	0.1171	0.1283	0.1188	0.1224
BOOSTLR	0.0897	0.0788	0.0957	0.1003	0.1009	0.1017	0.0880	0.0855	0.0926
GBM	0.0479	0.0460	0.0503	0.0506	0.0484	0.0477	0.0474	0.0489	0.0484
GBM-Stack	0.0463	0.0442	0.0498	0.0495	0.0481	0.0485	0.0478	0.0474	0.0477
KNN	0.1225	0.1322	0.1130	0.1310	0.1292	0.1436	0.1161	0.1122	0.1250
LR	0.0496	0.0432	0.0442	0.0562	0.0470	0.0502	0.0422	0.0462	0.0474
LR-Stack	0.0471	0.0442	0.0520	0.0526	0.0511	0.0511	0.0508	0.0486	0.0497
NB	0.0598	0.0537	0.0485	0.0669	0.0570	0.0586	0.0478	0.0565	0.0561
NNET	0.0469	0.0552	0.0651	0.0578	0.0555	0.0589	0.0582	0.0653	0.0579
RF	0.0498	0.0461	0.0524	0.0529	0.0505	0.0508	0.0477	0.0521	0.0503
RLR	0.0482	0.0423	0.0437	0.0537	0.0456	0.0484	0.0417	0.0453	0.0461
SC	0.0420	0.0382	0.0409	0.0424	0.0395	0.0406	0.0393	0.0413	0.0405
Superlearner	0.0457	0.0417	0.0442	0.0473	0.0443	0.0463	0.0425	0.0447	0.0446
SVM	0.0495	0.0439	0.0468	0.0522	0.0483	0.0501	0.0447	0.0475	0.0479
Twang-GBM	0.0447	0.0409	0.0441	0.0460	0.0427	0.0441	0.0427	0.0440	0.0437
Simulated PS	0.0470	0.0510	0.0556	0.0560	0.0551	0.0597	0.0536	0.0615	0.0550

Table 91: (MCse; n=1000; ATE= -0.062) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0479	0.0527	0.0576	0.0583	0.0573	0.0617	0.0519	0.0578	0.0557
BAG-CART	0.1158	0.1321	0.1370	0.1127	0.1430	0.1233	0.1325	0.1269	0.1279
BOOSTLR	0.0927	0.0808	0.1021	0.1049	0.1045	0.1057	0.0916	0.0910	0.0967
GBM	0.0496	0.0475	0.0531	0.0526	0.0509	0.0505	0.0502	0.0523	0.0508
GBM-Stack	0.0476	0.0459	0.0533	0.0512	0.0500	0.0513	0.0505	0.0505	0.0500
KNN	0.1264	0.1368	0.1208	0.1347	0.1351	0.1487	0.1234	0.1190	0.1306
LR	0.0513	0.0451	0.0473	0.0582	0.0501	0.0532	0.0445	0.0494	0.0499
LR-Stack	0.0488	0.0456	0.0551	0.0539	0.0533	0.0546	0.0531	0.0517	0.0520
NB	0.0624	0.0554	0.0516	0.0696	0.0603	0.0618	0.0507	0.0599	0.0589
NNET	0.0484	0.0569	0.0686	0.0588	0.0576	0.0614	0.0639	0.0691	0.0606
RF	0.0522	0.0477	0.0575	0.0550	0.0536	0.0538	0.0514	0.0556	0.0534
RLR	0.0499	0.0441	0.0467	0.0557	0.0485	0.0513	0.0439	0.0483	0.0486
SC	0.0430	0.0399	0.0433	0.0437	0.0414	0.0429	0.0412	0.0437	0.0424
Superlearner	0.0471	0.0432	0.0472	0.0490	0.0465	0.0493	0.0452	0.0478	0.0469
SVM	0.0503	0.0454	0.0499	0.0536	0.0503	0.0529	0.0474	0.0507	0.0501
Twang-GBM	0.0462	0.0424	0.0470	0.0479	0.0450	0.0471	0.0454	0.0473	0.0460
Simulated PS	0.0490	0.0526	0.0582	0.0581	0.0572	0.0622	0.0573	0.0656	0.0575

Table 92: (MCse; n=1000; ATE= -0.044) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0498	0.0546	0.0608	0.0620	0.0583	0.0647	0.0544	0.0608	0.0582
BAG-CART	0.1232	0.1356	0.1435	0.1203	0.1470	0.1283	0.1392	0.1326	0.1337
BOOSTLR	0.0967	0.0847	0.1059	0.1092	0.1078	0.1105	0.0946	0.0942	0.1004
GBM	0.0520	0.0495	0.0558	0.0558	0.0520	0.0520	0.0513	0.0552	0.0530
GBM-Stack	0.0497	0.0475	0.0558	0.0544	0.0512	0.0527	0.0521	0.0530	0.0520
KNN	0.1325	0.1414	0.1273	0.1425	0.1390	0.1534	0.1278	0.1254	0.1362
LR	0.0533	0.0473	0.0503	0.0622	0.0508	0.0551	0.0461	0.0526	0.0522
LR-Stack	0.0509	0.0473	0.0577	0.0571	0.0549	0.0560	0.0547	0.0544	0.0541
NB	0.0648	0.0579	0.0542	0.0756	0.0615	0.0644	0.0517	0.0629	0.0616
NNET	0.0504	0.0603	0.0720	0.0619	0.0591	0.0631	0.0666	0.0717	0.0631
RF	0.0549	0.0502	0.0602	0.0594	0.0555	0.0556	0.0536	0.0586	0.0560
RLR	0.0519	0.0462	0.0496	0.0593	0.0490	0.0531	0.0456	0.0514	0.0508
SC	0.0449	0.0418	0.0454	0.0462	0.0421	0.0441	0.0428	0.0459	0.0441
Superlearner	0.0491	0.0452	0.0497	0.0523	0.0475	0.0508	0.0467	0.0504	0.0490
SVM	0.0525	0.0472	0.0521	0.0567	0.0515	0.0542	0.0491	0.0531	0.0520
Twang-GBM	0.0485	0.0440	0.0493	0.0508	0.0461	0.0483	0.0469	0.0499	0.0480
Simulated PS	0.0510	0.0542	0.0608	0.0609	0.0576	0.0636	0.0593	0.0679	0.0594

Table 93: (MCse; n=1000; ATE= 0.055) Monte-Carlo standard error (MCse) of ATE estimate across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting MCse by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean MCse
AVNNET	0.0598	0.0628	0.0717	0.0802	0.0699	0.0780	0.0643	0.0738	0.0701
BAG-CART	0.1505	0.1570	0.1688	0.1585	0.1691	0.1604	0.1624	0.1590	0.1607
BOOSTLR	0.1178	0.1053	0.1221	0.1274	0.1238	0.1338	0.1087	0.1159	0.1194
GBM	0.0618	0.0590	0.0628	0.0689	0.0614	0.0635	0.0593	0.0641	0.0626
GBM-Stack	0.0590	0.0566	0.0638	0.0671	0.0592	0.0641	0.0607	0.0620	0.0616
KNN	0.1575	0.1636	0.1564	0.1659	0.1626	0.1742	0.1526	0.1505	0.1604
LR	0.0645	0.0584	0.0578	0.0812	0.0643	0.0698	0.0537	0.0649	0.0643
LR-Stack	0.0604	0.0565	0.0659	0.0713	0.0633	0.0678	0.0632	0.0634	0.0640
NB	0.0789	0.0682	0.0625	0.1020	0.0768	0.0825	0.0621	0.0773	0.0763
NNET	0.0608	0.0718	0.0848	0.0769	0.0732	0.0781	0.0814	0.0876	0.0768
RF	0.0679	0.0633	0.0701	0.0765	0.0673	0.0701	0.0644	0.0731	0.0691
RLR	0.0626	0.0570	0.0570	0.0769	0.0613	0.0671	0.0530	0.0631	0.0622
SC	0.0532	0.0509	0.0516	0.0561	0.0504	0.0541	0.0491	0.0546	0.0525
Superlearner	0.0585	0.0550	0.0574	0.0656	0.0568	0.0625	0.0549	0.0602	0.0589
SVM	0.0612	0.0565	0.0605	0.0684	0.0595	0.0649	0.0580	0.0624	0.0614
Twang-GBM	0.0573	0.0537	0.0566	0.0628	0.0549	0.0595	0.0547	0.0587	0.0573
Simulated PS	0.0618	0.0630	0.0705	0.0745	0.0660	0.0743	0.0702	0.0801	0.0700

Table 94: (SE; n=1000; ATE= -0.103) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.103 (corresponding to -1.2 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0416	0.0449	0.0449	0.0451	0.0459	0.0470	0.0440	0.0448	0.0448
BAG-CART	0.0559	0.0679	0.0689	0.0587	0.0680	0.0587	0.0659	0.0582	0.0628
BOOSTLR	0.0631	0.0610	0.0646	0.0655	0.0665	0.0692	0.0604	0.0583	0.0636
GBM	0.0420	0.0438	0.0439	0.0432	0.0437	0.0427	0.0436	0.0435	0.0433
GBM-Stack	0.0417	0.0431	0.0439	0.0437	0.0440	0.0439	0.0445	0.0428	0.0434
KNN	0.0714	0.0744	0.0689	0.0760	0.0775	0.0820	0.0700	0.0679	0.0735
LR	0.0429	0.0411	0.0407	0.0434	0.0409	0.0436	0.0402	0.0396	0.0415
LR-Stack	0.0423	0.0431	0.0452	0.0443	0.0439	0.0456	0.0463	0.0433	0.0443
NB	0.0472	0.0471	0.0421	0.0477	0.0462	0.0486	0.0427	0.0432	0.0456
NNET	0.0421	0.0477	0.0510	0.0461	0.0476	0.0480	0.0503	0.0490	0.0477
RF	0.0421	0.0429	0.0442	0.0433	0.0433	0.0427	0.0432	0.0428	0.0431
RLR	0.0423	0.0407	0.0403	0.0428	0.0404	0.0427	0.0398	0.0393	0.0410
SC	0.0386	0.0378	0.0380	0.0384	0.0378	0.0380	0.0378	0.0376	0.0380
Superlearner	0.0410	0.0408	0.0404	0.0415	0.0410	0.0417	0.0405	0.0399	0.0409
SVM	0.0442	0.0425	0.0420	0.0456	0.0439	0.0454	0.0427	0.0427	0.0436
Twang-GBM	0.0403	0.0405	0.0405	0.0408	0.0406	0.0405	0.0403	0.0400	0.0404
Simulated PS	0.0420	0.0450	0.0460	0.0466	0.0465	0.0481	0.0465	0.0469	0.0460

Table 95: (SE; n=1000; ATE= -0.091) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.091 (corresponding to -1.0 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0435	0.0464	0.0466	0.0474	0.0482	0.0487	0.0458	0.0468	0.0467
BAG-CART	0.0594	0.0696	0.0707	0.0615	0.0716	0.0619	0.0691	0.0622	0.0657
BOOSTLR	0.0656	0.0628	0.0675	0.0693	0.0694	0.0718	0.0632	0.0607	0.0663
GBM	0.0437	0.0451	0.0457	0.0451	0.0456	0.0443	0.0453	0.0452	0.0450
GBM-Stack	0.0435	0.0445	0.0456	0.0455	0.0458	0.0455	0.0464	0.0444	0.0451
KNN	0.0762	0.0780	0.0740	0.0798	0.0812	0.0856	0.0754	0.0715	0.0777
LR	0.0448	0.0426	0.0421	0.0456	0.0430	0.0455	0.0418	0.0413	0.0433
LR-Stack	0.0441	0.0445	0.0470	0.0463	0.0458	0.0473	0.0482	0.0449	0.0460
NB	0.0492	0.0485	0.0438	0.0501	0.0487	0.0507	0.0444	0.0453	0.0476
NNET	0.0440	0.0491	0.0530	0.0480	0.0496	0.0499	0.0525	0.0511	0.0497
RF	0.0443	0.0445	0.0459	0.0457	0.0455	0.0448	0.0451	0.0452	0.0451
RLR	0.0441	0.0421	0.0418	0.0449	0.0425	0.0445	0.0414	0.0409	0.0428
SC	0.0402	0.0391	0.0393	0.0401	0.0393	0.0394	0.0391	0.0392	0.0395
Superlearner	0.0429	0.0422	0.0420	0.0435	0.0428	0.0434	0.0421	0.0417	0.0426
SVM	0.0459	0.0439	0.0436	0.0475	0.0457	0.0469	0.0445	0.0443	0.0453
Twang-GBM	0.0421	0.0418	0.0421	0.0428	0.0425	0.0421	0.0420	0.0417	0.0421
Simulated PS	0.0437	0.0464	0.0477	0.0487	0.0484	0.0500	0.0485	0.0487	0.0478

Table 96: (SE; n=1000; ATE= -0.078) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.078 (corresponding to -0.8 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0454	0.0482	0.0488	0.0498	0.0502	0.0509	0.0480	0.0494	0.0488
BAG-CART	0.0626	0.0723	0.0743	0.0649	0.0765	0.0662	0.0731	0.0668	0.0696
BOOSTLR	0.0688	0.0656	0.0707	0.0725	0.0724	0.0749	0.0662	0.0634	0.0693
GBM	0.0455	0.0468	0.0477	0.0472	0.0475	0.0461	0.0470	0.0473	0.0469
GBM-Stack	0.0454	0.0461	0.0474	0.0475	0.0477	0.0473	0.0482	0.0465	0.0470
KNN	0.0817	0.0824	0.0782	0.0849	0.0872	0.0892	0.0800	0.0754	0.0824
LR	0.0467	0.0443	0.0439	0.0480	0.0454	0.0474	0.0435	0.0434	0.0453
LR-Stack	0.0460	0.0461	0.0489	0.0482	0.0477	0.0491	0.0502	0.0470	0.0479
NB	0.0513	0.0505	0.0456	0.0528	0.0512	0.0526	0.0462	0.0481	0.0498
NNET	0.0459	0.0511	0.0556	0.0503	0.0516	0.0518	0.0549	0.0539	0.0519
RF	0.0465	0.0463	0.0481	0.0481	0.0479	0.0470	0.0470	0.0477	0.0473
RLR	0.0460	0.0438	0.0435	0.0472	0.0448	0.0463	0.0431	0.0430	0.0447
SC	0.0420	0.0406	0.0408	0.0418	0.0410	0.0410	0.0406	0.0410	0.0411
Superlearner	0.0447	0.0438	0.0438	0.0456	0.0448	0.0452	0.0439	0.0438	0.0444
SVM	0.0477	0.0454	0.0455	0.0495	0.0475	0.0486	0.0463	0.0464	0.0471
Twang-GBM	0.0440	0.0435	0.0438	0.0448	0.0445	0.0439	0.0437	0.0438	0.0440
Simulated PS	0.0457	0.0481	0.0500	0.0509	0.0505	0.0518	0.0505	0.0511	0.0498

Table 97: (SE; n=1000; ATE= -0.062) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.062 (corresponding to -0.6 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0477	0.0502	0.0508	0.0522	0.0526	0.0534	0.0502	0.0520	0.0511
BAG-CART	0.0663	0.0760	0.0775	0.0690	0.0809	0.0700	0.0763	0.0718	0.0735
BOOSTLR	0.0720	0.0689	0.0748	0.0760	0.0760	0.0779	0.0694	0.0672	0.0728
GBM	0.0478	0.0485	0.0497	0.0494	0.0495	0.0482	0.0492	0.0496	0.0490
GBM-Stack	0.0475	0.0478	0.0494	0.0498	0.0495	0.0493	0.0503	0.0488	0.0491
KNN	0.0866	0.0872	0.0843	0.0900	0.0928	0.0949	0.0855	0.0804	0.0877
LR	0.0493	0.0463	0.0459	0.0504	0.0476	0.0496	0.0454	0.0456	0.0475
LR-Stack	0.0483	0.0478	0.0509	0.0505	0.0496	0.0513	0.0523	0.0493	0.0500
NB	0.0545	0.0525	0.0476	0.0556	0.0536	0.0549	0.0482	0.0509	0.0522
NNET	0.0484	0.0531	0.0582	0.0525	0.0538	0.0542	0.0580	0.0571	0.0544
RF	0.0491	0.0483	0.0504	0.0506	0.0503	0.0493	0.0492	0.0507	0.0497
RLR	0.0485	0.0457	0.0454	0.0496	0.0468	0.0484	0.0449	0.0451	0.0468
SC	0.0439	0.0423	0.0424	0.0437	0.0427	0.0427	0.0422	0.0430	0.0429
Superlearner	0.0470	0.0456	0.0457	0.0478	0.0467	0.0472	0.0458	0.0461	0.0465
SVM	0.0497	0.0471	0.0474	0.0515	0.0493	0.0506	0.0483	0.0485	0.0491
Twang-GBM	0.0461	0.0453	0.0457	0.0470	0.0464	0.0459	0.0456	0.0461	0.0460
Simulated PS	0.0482	0.0500	0.0524	0.0534	0.0525	0.0542	0.0529	0.0539	0.0522

Table 98: (SE; n=1000; ATE= -0.044) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = -0.044 (corresponding to -0.4 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0500	0.0523	0.0531	0.0552	0.0550	0.0561	0.0527	0.0547	0.0536
BAG-CART	0.0707	0.0799	0.0816	0.0744	0.0853	0.0742	0.0805	0.0776	0.0780
BOOSTLR	0.0755	0.0726	0.0786	0.0803	0.0800	0.0816	0.0729	0.0704	0.0765
GBM	0.0501	0.0506	0.0519	0.0522	0.0517	0.0503	0.0512	0.0521	0.0513
GBM-Stack	0.0498	0.0497	0.0517	0.0523	0.0517	0.0513	0.0523	0.0512	0.0513
KNN	0.0931	0.0924	0.0908	0.0967	0.0981	0.0997	0.0908	0.0859	0.0934
LR	0.0518	0.0485	0.0479	0.0540	0.0500	0.0519	0.0472	0.0480	0.0499
LR-Stack	0.0506	0.0497	0.0533	0.0530	0.0517	0.0534	0.0544	0.0516	0.0522
NB	0.0572	0.0551	0.0498	0.0605	0.0564	0.0575	0.0503	0.0540	0.0551
NNET	0.0508	0.0554	0.0610	0.0558	0.0563	0.0568	0.0609	0.0599	0.0571
RF	0.0519	0.0507	0.0530	0.0539	0.0528	0.0517	0.0515	0.0533	0.0524
RLR	0.0509	0.0478	0.0474	0.0530	0.0492	0.0507	0.0467	0.0475	0.0491
SC	0.0460	0.0440	0.0441	0.0459	0.0446	0.0445	0.0438	0.0450	0.0447
Superlearner	0.0492	0.0475	0.0477	0.0505	0.0489	0.0493	0.0477	0.0484	0.0487
SVM	0.0519	0.0490	0.0495	0.0539	0.0515	0.0526	0.0502	0.0509	0.0512
Twang-GBM	0.0484	0.0473	0.0478	0.0495	0.0485	0.0480	0.0476	0.0486	0.0482
Simulated PS	0.0506	0.0521	0.0549	0.0562	0.0548	0.0564	0.0555	0.0567	0.0547

Table 99: (SE; n=1000; ATE= 0.055) Estimated robust sandwich-type standard error (SE) of ATE estimate averaged across the m=1000 simulated data sets with true (marginal risk difference) ATE = 0.055 (corresponding to 0.4 in log-odds scale), based on the n=1000 simulated data setup. Models are ordered alphabetically. The last row (Simulated PS) presents the resulting SE by using the actually simulated propensity scores, without estimating them by any model.

	A	B	C	D	E	F	G	H	Mean SE
AVNNET	0.0605	0.0618	0.0635	0.0684	0.0658	0.0681	0.0622	0.0680	0.0648
BAG-CART	0.0932	0.1002	0.1043	0.0991	0.1078	0.0978	0.1016	0.1024	0.1008
BOOSTLR	0.0923	0.0891	0.0963	0.0996	0.0967	0.1006	0.0888	0.0879	0.0939
GBM	0.0605	0.0594	0.0611	0.0636	0.0617	0.0605	0.0603	0.0635	0.0613
GBM-Stack	0.0600	0.0582	0.0608	0.0629	0.0611	0.0612	0.0609	0.0616	0.0609
KNN	0.1204	0.1183	0.1173	0.1231	0.1248	0.1275	0.1192	0.1139	0.1206
LR	0.0634	0.0582	0.0564	0.0690	0.0628	0.0634	0.0549	0.0606	0.0611
LR-Stack	0.0608	0.0583	0.0628	0.0639	0.0614	0.0633	0.0634	0.0620	0.0620
NB	0.0721	0.0653	0.0593	0.0800	0.0708	0.0713	0.0594	0.0699	0.0685
NNET	0.0617	0.0664	0.0740	0.0678	0.0672	0.0691	0.0731	0.0747	0.0692
RF	0.0649	0.0616	0.0637	0.0676	0.0652	0.0644	0.0619	0.0668	0.0645
RLR	0.0621	0.0572	0.0557	0.0671	0.0611	0.0616	0.0543	0.0594	0.0598
SC	0.0548	0.0515	0.0515	0.0549	0.0527	0.0526	0.0509	0.0540	0.0529
Superlearner	0.0595	0.0561	0.0563	0.0617	0.0586	0.0593	0.0559	0.0591	0.0583
SVM	0.0612	0.0573	0.0585	0.0641	0.0608	0.0621	0.0588	0.0609	0.0605
Twang-GBM	0.0583	0.0558	0.0566	0.0601	0.0579	0.0575	0.0559	0.0591	0.0576
Simulated PS	0.0616	0.0612	0.0654	0.0685	0.0652	0.0680	0.0659	0.0693	0.0656

Table 100: (Raw (naive) bias; n=1000) Summary table of relative ATE estimation bias (in %) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Relative bias is presented across all eight simulated scenarios (A-H) and across all six simulated average treatment effects based on the n=1000 data setup.

	A	B	C	D	E	F	G	H	Mean-bias
-1.2	16.88	20.57	3.98	23.46	23.17	21.43	5.33	17.51	16.54
-1.0	20.87	24.71	4.58	28.59	28.43	25.76	5.76	21.58	20.03
-0.8	26.91	31.10	5.89	35.92	35.19	31.94	6.52	28.05	25.19
-0.6	36.13	41.59	7.35	48.52	47.52	42.43	7.93	38.69	33.77
-0.4	56.04	62.79	10.85	75.29	73.12	64.36	11.82	60.91	51.90
+0.4	59.70	61.79	13.71	77.74	75.59	69.45	15.06	65.73	54.85

Table 101: (n=1000) Summary table of ATE estimation mean-squared error (mse) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. mean-squared error (mse) computed across m=1000 simulated data set based on the n=1000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-mse
-1.2	0.0018	0.0018	0.0014	0.0021	0.0020	0.0019	0.0014	0.0018	0.0018
-1.0	0.0020	0.0019	0.0016	0.0024	0.0022	0.0021	0.0014	0.0020	0.0019
-0.8	0.0022	0.0021	0.0017	0.0026	0.0023	0.0023	0.0015	0.0023	0.0021
-0.6	0.0024	0.0023	0.0018	0.0029	0.0026	0.0026	0.0017	0.0025	0.0023
-0.4	0.0027	0.0025	0.0020	0.0033	0.0029	0.0028	0.0018	0.0029	0.0026
+0.4	0.0040	0.0038	0.0026	0.0050	0.0044	0.0044	0.0024	0.0044	0.0039

Table 102: (n=1000) Summary table of ATE estimation Monte Carlo standard error (MCse) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Monte Carlo standard error (MCse) computed across m=1000 simulated data set based on the n=1000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-MCse
-1.2	0.0381	0.0365	0.0377	0.0394	0.0377	0.0376	0.0374	0.0391	0.0379
-1.0	0.0402	0.0381	0.0393	0.0409	0.0392	0.0390	0.0379	0.0404	0.0394
-0.8	0.0423	0.0385	0.0405	0.0431	0.0396	0.0409	0.0389	0.0421	0.0407
-0.6	0.0435	0.0403	0.0428	0.0446	0.0414	0.0430	0.0409	0.0441	0.0425
-0.4	0.0455	0.0421	0.0447	0.0472	0.0425	0.0443	0.0423	0.0463	0.0444
+0.4	0.0539	0.0513	0.0507	0.0557	0.0508	0.0535	0.0485	0.0548	0.0524

Table 103: (n=1000) Summary table of ATE estimation Robust sandwich-type standard error (SE) when ignoring selection bias and using a naive treatment effect estimate without propensity score based IPTW. Robust sandwich-type standard error (SE) computed across m=1000 simulated data set based on the n=1000 simulated data setup, and presented for all eight simulated scenarios (A-H) and all six simulated average treatment effects.

	A	B	C	D	E	F	G	H	Mean-bias
-1.2	0.0386	0.0378	0.0377	0.0388	0.0382	0.0380	0.0374	0.0382	0.0381
-1.0	0.0403	0.0392	0.0390	0.0405	0.0398	0.0395	0.0388	0.0400	0.0396
-0.8	0.0421	0.0408	0.0405	0.0423	0.0414	0.0411	0.0402	0.0419	0.0413
-0.6	0.0441	0.0425	0.0421	0.0443	0.0432	0.0429	0.0418	0.0440	0.0431
-0.4	0.0461	0.0443	0.0437	0.0465	0.0451	0.0448	0.0434	0.0461	0.0450
+0.4	0.0545	0.0515	0.0510	0.0548	0.0527	0.0526	0.0504	0.0547	0.0528

Table 104: (ASAM; n=1000) Mean of the averaged standardized absolute mean differences over m=1000 simulated data sets in the n=1000 simulated data setup. In each data set the mean of the standardized absolute mean differences of all ten covariates is taken. We describe values in this table as ASAM. The last row (NO WEIGHT) presents the ASAM in the initial non-weighted data.

	A	B	C	D	E	F	G	H	Mean-ASAM
AVNNET	0.150	0.151	0.151	0.181	0.163	0.176	0.147	0.175	0.162
BAG-CART	0.338	0.355	0.358	0.386	0.369	0.355	0.334	0.348	0.355
BOOSTLR	0.375	0.354	0.371	0.441	0.424	0.459	0.339	0.376	0.392
GBM	0.156	0.148	0.150	0.164	0.156	0.153	0.142	0.158	0.153
GBM-Stack	0.151	0.143	0.148	0.158	0.150	0.149	0.145	0.149	0.149
KNN	0.413	0.458	0.364	0.452	0.466	0.483	0.390	0.414	0.430
LR	0.157	0.158	0.136	0.210	0.188	0.164	0.130	0.173	0.164
LR-Stack	0.161	0.152	0.164	0.170	0.159	0.167	0.164	0.159	0.162
NB	0.262	0.204	0.154	0.396	0.273	0.263	0.156	0.239	0.243
NNET	0.151	0.172	0.191	0.181	0.175	0.180	0.193	0.208	0.182
RF	0.161	0.147	0.153	0.170	0.157	0.157	0.142	0.162	0.156
RLR	0.153	0.151	0.133	0.194	0.174	0.156	0.128	0.164	0.157
SC	0.206	0.180	0.143	0.220	0.203	0.223	0.149	0.174	0.187
Superlearner	0.147	0.137	0.133	0.153	0.146	0.145	0.129	0.142	0.142
SVM	0.160	0.141	0.139	0.166	0.150	0.155	0.140	0.150	0.150
Twang-GBM	0.144	0.137	0.122	0.153	0.149	0.150	0.118	0.140	0.139
Simulated PS	0.136	0.130	0.144	0.156	0.143	0.155	0.152	0.158	0.147
NO WEIGHT	0.299	0.274	0.170	0.349	0.320	0.332	0.177	0.270	0.274

Table 105: (ASAM$_{conf}$; n=1000) Mean of the averaged standardized absolute mean differences over m=1000 simulated data sets in the n=1000 simulated data setup. In each data set, the mean of the standardized absolute mean differences of the four confounding covariates is taken. We therefore describe values in this table with ASAM$_{conf}$. The last row (NO WEIGHT) presents the ASAM$_{conf}$ in the initial non-weighted data.

	A	B	C	D	E	F	G	H	Mean-ASAM
AVN	0.150	0.152	0.153	0.183	0.164	0.175	0.151	0.177	0.163
BAG-CART	0.337	0.355	0.348	0.389	0.375	0.365	0.328	0.351	0.356
BOOSTLR	0.375	0.351	0.404	0.522	0.506	0.548	0.374	0.418	0.437
GBM	0.160	0.149	0.157	0.174	0.162	0.162	0.151	0.163	0.160
GBM-Stack	0.150	0.141	0.149	0.162	0.154	0.156	0.155	0.154	0.152
KNN	0.397	0.458	0.362	0.436	0.471	0.479	0.391	0.413	0.426
LR	0.158	0.168	0.132	0.223	0.204	0.168	0.132	0.188	0.172
LR-Stack	0.160	0.149	0.161	0.176	0.161	0.176	0.173	0.164	0.165
NB	0.266	0.205	0.153	0.424	0.285	0.269	0.158	0.248	0.251
NNET	0.151	0.173	0.197	0.184	0.178	0.183	0.197	0.214	0.185
RF	0.159	0.142	0.160	0.170	0.156	0.166	0.149	0.162	0.158
RLR	0.154	0.160	0.130	0.205	0.187	0.161	0.130	0.176	0.163
SC	0.211	0.180	0.151	0.264	0.238	0.308	0.175	0.203	0.216
Superlearner	0.148	0.136	0.136	0.159	0.150	0.154	0.138	0.147	0.146
SVM	0.163	0.140	0.141	0.166	0.151	0.161	0.145	0.152	0.152
Twang-GBM	0.147	0.138	0.126	0.166	0.159	0.176	0.132	0.150	0.149
Simulated PS	0.137	0.131	0.149	0.158	0.144	0.158	0.157	0.160	0.149
NO WEIGHT	0.318	0.290	0.175	0.428	0.389	0.458	0.219	0.339	0.327

Table 106: Average of maximum IPTW weights resulting through the different propensity score estimation models in each of the m=1000 simulated data sets for each scenario (A-H) in the n=1000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVNNET	11.7	16.4	18.0	26.5	19.8	28.8	16.0	28.5	20.7
BAG-CART	150.2	146.3	149.7	169.8	158.3	154.2	143.6	154.6	153.3
BOOSTLR	79.7	74.9	94.8	116.0	102.9	117.2	73.6	79.9	92.4
GBM	13.7	12.1	13.5	16.9	14.4	13.2	12.5	16.1	14.1
GBM-Stack	9.9	9.0	10.4	11.1	10.2	10.4	10.6	10.4	10.2
KNN	197.2	195.0	191.8	198.5	197.9	196.8	190.9	197.7	195.7
LR	18.9	18.3	9.3	41.6	30.6	23.5	6.8	23.7	21.6
LR-Stack	12.3	9.4	13.6	14.5	11.7	15.7	14.4	11.4	12.9
NB	50.1	26.4	14.8	107.7	48.7	46.4	16.5	81.3	49.0
NNET	13.5	23.4	35.1	24.1	23.6	30.0	36.0	41.1	28.4
RF	28.2	20.8	22.0	31.4	24.7	24.3	19.0	30.6	25.1
RLR	16.5	15.9	8.5	34.0	25.4	19.8	6.2	20.2	18.3
SC	5.8	5.2	3.7	7.5	6.5	5.4	3.2	6.7	5.5
Superlearner	8.6	7.5	7.5	9.9	8.6	8.7	7.3	8.5	8.3
SVM	10.9	8.8	9.8	13.0	11.1	12.2	10.0	11.2	10.9
Twang-GBM	11.4	9.1	8.2	14.5	11.3	12.3	8.0	11.7	10.8
Simulated PS	16.3	15.2	21.4	21.8	18.0	23.3	24.2	23.7	20.5

Table 107: Average of mean IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=1000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVNNET	2.05	2.13	2.09	2.20	2.14	2.22	2.05	2.23	2.14
BAG-CART	3.39	3.40	3.39	3.76	3.58	3.38	3.26	3.49	3.46
BOOSTLR	3.69	3.60	3.66	4.38	4.14	4.29	3.39	3.65	3.85
GBM	2.03	2.03	2.00	2.04	2.04	2.00	2.01	2.08	2.03
GBM-Stack	2.03	2.04	2.03	2.04	2.04	2.04	2.04	2.05	2.04
KNN	6.79	6.50	7.53	6.89	6.66	6.42	6.70	7.12	6.83
LR	2.06	2.10	2.06	2.21	2.16	2.09	2.04	2.13	2.10
LR-Stack	2.07	2.07	2.08	2.10	2.10	2.17	2.12	2.07	2.10
NB	2.63	2.34	1.99	3.19	2.65	2.47	2.06	3.62	2.62
NNET	2.05	2.25	2.37	2.17	2.21	2.24	2.35	2.44	2.26
RF	2.21	2.14	2.12	2.16	2.12	2.10	2.07	2.18	2.14
RLR	2.02	2.06	2.04	2.13	2.10	2.04	2.02	2.08	2.06
SC	1.91	1.93	1.98	1.89	1.90	1.90	1.98	1.94	1.93
Superlearner	1.74	1.74	1.71	1.70	1.72	1.71	1.71	1.71	1.72
SVM	2.03	2.01	1.99	2.01	2.01	2.01	1.99	2.00	2.01
Twang-GBM	1.99	1.97	1.91	1.97	1.96	1.96	1.91	1.98	1.96
Simulated PS	2.00	2.00	1.99	2.00	2.00	2.00	1.99	1.99	2.00

Table 108: Average of the first quantile of the IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=1000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVN	1.19	1.24	1.24	1.15	1.18	1.17	1.24	1.17	1.20
BAG-CART	1.21	1.23	1.20	1.14	1.18	1.18	1.21	1.16	1.19
BOOSTLR	1.20	1.27	1.23	1.11	1.16	1.11	1.28	1.19	1.20
GBM	1.23	1.26	1.25	1.17	1.21	1.22	1.27	1.19	1.23
GBM-Stack	1.21	1.26	1.23	1.16	1.20	1.20	1.23	1.19	1.21
KNN	1.15	1.20	1.21	1.09	1.13	1.12	1.20	1.12	1.15
LR	1.20	1.31	1.44	1.16	1.23	1.21	1.47	1.25	1.28
LR-Stack	1.21	1.25	1.21	1.16	1.20	1.19	1.20	1.19	1.20
NB	1.13	1.18	1.26	1.10	1.13	1.15	1.27	1.07	1.16
NNET	1.19	1.22	1.20	1.15	1.18	1.18	1.19	1.16	1.18
RF	1.21	1.26	1.26	1.17	1.21	1.21	1.28	1.18	1.22
RLR	1.22	1.33	1.46	1.18	1.24	1.23	1.49	1.26	1.30
SC	1.34	1.51	1.57	1.33	1.42	1.42	1.62	1.36	1.45
SL	1.19	1.23	1.22	1.15	1.19	1.19	1.24	1.18	1.20
SVM	1.25	1.28	1.27	1.18	1.21	1.20	1.26	1.22	1.23
Twang-GBM	1.23	1.29	1.31	1.18	1.23	1.23	1.32	1.22	1.25
Simulated PS	1.21	1.24	1.20	1.13	1.17	1.15	1.18	1.13	1.18

Table 109: Average of the third quantile of the IPTW weights over m=1000 simulations resulting through the different propensity score estimation models for each scenario (A-H) in the n=1000 simulated data setup.

	A	B	C	D	E	F	G	H	Average
AVN	2.14	2.23	2.18	1.98	2.08	2.09	2.16	2.05	2.11
BAG-CART	2.30	2.34	2.27	2.10	2.21	2.18	2.27	2.12	2.22
BOOSTLR	3.69	3.72	3.67	3.51	3.68	3.67	3.71	3.64	3.66
GBM	2.15	2.19	2.13	1.99	2.09	2.08	2.16	2.06	2.11
GBM-Stack	2.16	2.24	2.16	1.97	2.10	2.10	2.17	2.06	2.12
KNN	2.44	2.97	6.17	2.40	2.35	2.32	4.35	2.72	3.21
LR	2.11	2.18	2.27	1.98	2.07	2.10	2.31	2.10	2.14
LR-Stack	2.18	2.24	2.16	1.98	2.10	2.14	2.18	2.06	2.13
NB	2.22	2.26	2.14	2.04	2.13	2.19	2.22	2.34	2.19
NNET	2.13	2.27	2.28	1.98	2.11	2.14	2.25	2.14	2.16
RF	2.18	2.21	2.18	1.99	2.08	2.08	2.16	2.04	2.12
RLR	2.11	2.17	2.25	1.98	2.06	2.10	2.29	2.09	2.13
SC	2.27	2.13	2.39	2.14	2.10	2.16	2.35	2.31	2.23
SL	1.85	1.87	1.82	1.72	1.78	1.79	1.85	1.74	1.80
SVM	2.20	2.22	2.17	1.95	2.07	2.06	2.16	2.07	2.11
Twang-GBM	2.12	2.16	2.10	1.95	2.04	2.05	2.10	2.03	2.07
Simulated PS	2.09	2.11	2.04	1.89	1.97	1.95	2.02	1.87	1.99

12.4. Appendix A – Additional simulation study result tables

Table 110: Mean of the averaged standardized absolute mean differences over m=1000 simulated data sets, assuming that the control and treatment groups were received through random sampling. The values present the ASAM that would results from a randomized control trial based on our simulated covariates, for the three different data sizes, respectively.

Data setup	n = 1000	n= 2000	n = 3000
RCT ASAM	0.0612	0.077	0.1067

Table 111: Relative bias (in %) summarized for each scenario A-H. Results presented are averaged across the simulated ATEs $\gamma_{LO} \in \{-1.2, -1.0, -0.8, -0.6, -0.4, +0.4\}$ and data sizes $n \in \{3000, 2000, 1000\}$ for each scenario, respectively. The average across scenarios is shown in the last column.

	A	B	C	D	E	F	G	H	Mean bias
GBM-Stack	0.38	0.95	1.81	1.06	1.52	2.58	2.14	1.73	1.52
LR-Stack	1.50	1.17	1.41	1.97	1.52	2.13	1.96	2.16	1.73
SL	2.05	2.64	2.79	3.32	4.80	5.16	1.74	3.17	3.21
AVN	2.78	2.18	1.35	1.75	3.32	1.05	2.81	3.17	2.30
NNET	1.03	1.64	1.80	2.17	1.58	4.22	5.11	6.86	3.05
RLR	0.54	2.50	3.80	5.80	4.95	1.25	6.08	6.59	3.94
LR	0.84	2.42	4.66	7.80	6.34	1.54	6.87	8.13	4.83
Twang-GBM	8.60	10.46	4.50	13.15	13.20	12.00	2.52	9.88	9.29

Table 112: Ratio of the MC-SE and estimated SE (MC-SE / SE) for the different ATE estimators. The ratios are computed for each setup and then averaged across scenarios (A-H) and data sizes $n \in \{3000, 2000, 1000\}$, respectively. A ratio greater than one means that the estimated SE (Lumley, 2004) underestimated the empirical Monte-Carlo standard errors (MC-SE).

	-1.2	-1.0	-0.8	-0.6	-0.4	0.4	Average
GBM-Stack	1.023	1.022	1.020	1.020	1.019	1.013	1.019
LR-Stack	1.036	1.033	1.030	1.029	1.029	1.024	1.030
SL	1.017	1.016	1.013	1.014	1.014	1.010	1.014
AVN	1.058	1.055	1.053	1.051	1.050	1.042	1.051
NNET	1.096	1.091	1.087	1.087	1.082	1.074	1.086
RLR	1.033	1.035	1.036	1.037	1.037	1.032	1.035
LR	1.040	1.041	1.043	1.044	1.044	1.038	1.042
Twang-GBM	1.006	1.003	1.000	1.001	1.000	0.993	1.001

12.5. Appendix A – Correlations of ASAM and ATE estimates

Table 113: (n=3000) Pearson correlations of ASAM and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET and Twang-GBM, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.04	0.02	0.06	0.06	-0.01	0.03	0.10	0.01	0.04
-1.0	0.03	0.02	0.06	0.07	-0.01	0.02	0.09	0.03	0.04
-0.8	0.04	0.03	0.06	0.09	0.03	0.04	0.10	0.03	0.05
-0.6	0.04	0.02	0.05	0.09	0.05	0.05	0.10	0.05	0.06
-0.4	0.03	0.05	0.06	0.11	0.06	0.06	0.11	0.04	0.07
+0.4	0.02	0.07	0.07	0.17	0.13	0.05	0.10	0.09	0.09

Table 114: (n=3000) Spearman correlations of ASAM and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET and Twang-GBM, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.03	0.02	0.06	0.05	0.01	0.03	0.09	0.01	0.04
-1.0	0.02	0.02	0.06	0.07	0.01	0.02	0.08	0.02	0.04
-0.8	0.01	0.04	0.05	0.08	0.03	0.04	0.10	0.03	0.05
-0.6	0.03	0.02	0.05	0.07	0.04	0.04	0.10	0.05	0.05
-0.4	0.03	0.04	0.05	0.08	0.04	0.05	0.10	0.04	0.05
+0.4	0.02	0.07	0.06	0.12	0.09	0.03	0.09	0.07	0.07

Table 115: (n=3000) Pearson correlations of ASAM$_{conf}$ and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM$_{conf}$ (confounders only) from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET, and Twang-GBM, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.04	0.01	0.05	0.04	-0.03	-0.01	0.08	-0.01	0.02
-1.0	0.04	0.01	0.04	0.06	-0.03	-0.01	0.07	0.00	0.02
-0.8	0.04	0.01	0.05	0.07	0.02	0.00	0.09	0.02	0.04
-0.6	0.03	0.01	0.05	0.07	0.04	0.00	0.07	0.02	0.04
-0.4	0.02	0.04	0.06	0.08	0.04	0.03	0.07	0.03	0.05
+0.4	0.01	0.06	0.06	0.15	0.13	0.00	0.06	0.06	0.07

Table 116: (n=3000) Pearson correlations of ASAM and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET, Twang-GBM and the two models KNN and BAG-CART, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.46	0.52	0.45	0.50	0.53	0.53	0.50	0.43	0.49
-1.0	0.49	0.52	0.45	0.49	0.52	0.52	0.50	0.44	0.49
-0.8	0.51	0.53	0.45	0.49	0.53	0.52	0.51	0.46	0.50
-0.6	0.53	0.52	0.47	0.51	0.54	0.53	0.51	0.48	0.51
-0.4	0.54	0.55	0.47	0.51	0.55	0.54	0.51	0.50	0.52
+0.4	0.54	0.56	0.54	0.52	0.55	0.55	0.53	0.49	0.54

Table 117: (n=3000) Pearson correlations between the ASAM values and the absolute ATE estimation errors over m=1000 simulations of each model within each scenario (A-H).

	A	B	C	D	E	F	G	H	Mean
AVNNET	0.04	0.07	0.02	0.07	0.05	0.08	0.01	0.08	0.05
BAG-CART	0.18	0.26	0.09	0.18	0.23	0.22	0.18	0.23	0.20
BOOSTLR	0.27	0.21	0.21	0.24	0.23	0.23	0.27	0.27	0.24
GBM	-0.00	0.03	0.01	-0.02	-0.03	-0.01	0.04	0.08	0.01
GBM-Stack	0.02	-0.00	0.03	-0.02	-0.02	0.02	0.01	0.07	0.01
KNN	0.14	0.16	0.13	0.18	0.17	0.09	0.10	0.11	0.13
LR	0.04	0.00	-0.00	0.14	0.11	0.12	0.00	0.08	0.06
LR-Stack	0.10	0.02	0.01	0.01	-0.02	0.09	0.01	0.06	0.04
NB	0.21	0.16	0.14	0.21	0.23	0.23	0.15	0.13	0.18
NNET	0.06	0.07	0.11	0.16	0.12	0.10	0.14	0.14	0.11
RF	0.10	0.07	0.07	0.09	0.10	0.08	0.05	0.13	0.09
RLR	0.04	-0.00	-0.00	0.12	0.11	0.10	0.01	0.08	0.06
SC	0.01	0.07	0.06	0.02	0.03	0.06	-0.00	0.07	0.04
SL	0.01	0.01	0.00	-0.01	-0.03	-0.01	-0.01	0.06	0.00
SVM	0.06	0.01	-0.03	0.02	0.03	-0.02	0.05	0.10	0.03
Twang-GBM	0.00	0.07	0.00	0.00	-0.02	0.03	-0.03	0.07	0.02
True PS	0.02	0.23	0.11	0.15	0.30	0.14	0.15	0.21	0.16

Table 118: (n=2000) Pearson correlations of ASAM and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET and Twang-GBM, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.03	0.01	0.06	0.04	-0.02	0.06	0.05	0.01	0.03
-1.0	0.02	0.02	0.05	0.07	-0.01	0.05	0.04	0.03	0.03
-0.8	0.01	0.02	0.04	0.08	0.01	0.07	0.04	0.04	0.04
-0.6	0.01	0.03	0.04	0.12	0.03	0.08	0.03	0.04	0.05
-0.4	0.04	0.05	0.04	0.15	0.07	0.07	0.04	0.05	0.06
0.4	-0.00	0.08	0.03	0.16	0.08	0.09	0.02	0.08	0.07

Table 119: (n=2000) Pearson correlations of ASAM$_{conf}$ and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM$_{conf}$ (confounders only) from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET, and Twang-GBM, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.04	0.02	0.06	0.03	-0.04	0.03	0.05	0.01	0.02
-1.0	0.03	0.01	0.05	0.05	-0.02	0.02	0.04	0.02	0.03
-0.8	0.02	0.02	0.04	0.06	0.01	0.05	0.05	0.03	0.03
-0.6	0.02	0.03	0.03	0.10	0.03	0.05	0.05	0.03	0.04
-0.4	0.02	0.05	0.05	0.13	0.06	0.04	0.04	0.05	0.06
+0.4	0.01	0.08	0.04	0.13	0.10	0.05	0.03	0.07	0.07

Table 120: (n=2000) Pearson correlations of ASAM and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET, Twang-GBM and the two models KNN and BAG-CART, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.41	0.49	0.40	0.45	0.46	0.48	0.43	0.41	0.44
-1.0	0.44	0.49	0.42	0.43	0.46	0.47	0.43	0.44	0.45
-0.8	0.47	0.49	0.42	0.44	0.47	0.48	0.43	0.43	0.45
-0.6	0.51	0.50	0.44	0.46	0.48	0.50	0.44	0.45	0.47
-0.4	0.52	0.52	0.45	0.48	0.49	0.51	0.45	0.46	0.48
+0.4	0.55	0.58	0.52	0.53	0.54	0.53	0.53	0.50	0.53

Table 121: (n=2000) Pearson correlations between the ASAM values and the absolute ATE estimation errors over m=1000 simulations of each model within each scenario (A-H).

	A	B	C	D	E	F	G	H	Mean
AVNNET	-0.07	0.05	0.07	0.05	0.07	0.11	0.04	0.06	0.05
BAG-CART	0.19	0.28	0.15	0.19	0.25	0.20	0.15	0.25	0.21
BOOSTLR	0.23	0.17	0.17	0.19	0.25	0.19	0.18	0.27	0.21
GBM	-0.08	0.01	0.07	0.01	0.03	0.07	0.05	0.03	0.03
GBM-Stack	-0.05	0.02	0.05	0.01	0.01	0.07	0.00	0.05	0.02
KNN	0.16	0.18	0.10	0.21	0.16	0.12	0.14	0.21	0.16
LR	-0.05	0.01	0.02	0.16	0.05	0.11	0.03	0.10	0.06
LR-Stack	0.02	0.01	0.07	0.09	-0.00	0.16	-0.02	0.06	0.05
NB	0.17	0.24	0.07	0.27	0.17	0.23	0.22	0.25	0.20
NNET	-0.05	0.12	0.12	0.07	0.17	0.19	0.14	0.12	0.11
RF	-0.01	0.08	0.14	0.04	0.05	0.11	0.08	0.07	0.07
RLR	-0.06	0.01	0.02	0.14	0.04	0.10	0.03	0.09	0.05
SC	0.02	0.06	-0.01	0.05	0.06	0.09	0.00	0.01	0.03
SL	-0.05	0.00	0.05	0.02	0.00	0.06	0.00	0.03	0.01
SVM	-0.03	-0.00	0.03	0.02	0.03	0.08	0.01	0.04	0.02
Twang-GBM	-0.04	0.03	0.03	0.02	-0.02	0.05	0.00	0.03	0.01
True PS	-0.06	0.03	0.04	-0.02	-0.04	0.01	0.01	0.01	-0.00

Table 122: (n=2000) Average of the number of variables with an ASAM greater than 0.2 over m=1000 simulations in each Scenario (A-H). .

	A	B	C	D	E	F	G	H	Mean
GBM-Stack	1.270	1.173	1.373	1.609	1.345	1.360	1.286	1.338	1.344
LR-Stack	1.489	1.366	1.756	1.908	1.441	1.730	1.693	1.533	1.615
True PS	0.841	0.932	1.356	1.578	1.322	1.471	1.506	1.573	1.322
LR	1.348	1.597	0.976	2.994	2.559	1.736	0.816	2.109	1.767
NNET	1.276	1.526	1.890	2.280	1.897	2.409	1.926	2.493	1.962
AVNNET	1.282	1.219	1.208	2.100	1.489	1.772	1.198	1.754	1.503

Table 123: (n=1000) Pearson correlations of ASAM and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET and Twang-GBM, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.04	0.07	0.11	0.08	0.04	0.06	0.07	0.07	0.07
-1.0	0.04	0.07	0.11	0.12	0.04	0.09	0.10	0.06	0.08
-0.8	0.05	0.08	0.10	0.12	0.06	0.08	0.12	0.08	0.09
-0.6	0.06	0.07	0.10	0.14	0.09	0.09	0.11	0.09	0.09
-0.4	0.04	0.07	0.13	0.14	0.08	0.09	0.11	0.11	0.10
+0.4	0.05	0.11	0.12	0.20	0.13	0.10	0.10	0.14	0.12

Table 124: (n=1000) Pearson correlations of ASAM$_{conf}$ and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM$_{conf}$ (confounders only) from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET, and Twang-GBM, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.02	0.04	0.09	0.06	0.01	0.02	0.07	0.03	0.04
-1.0	0.03	0.04	0.09	0.09	0.01	0.05	0.10	0.03	0.05
-0.8	0.04	0.06	0.08	0.10	0.04	0.04	0.10	0.04	0.06
-0.6	0.04	0.04	0.10	0.12	0.06	0.04	0.09	0.05	0.07
-0.4	0.05	0.03	0.12	0.13	0.05	0.03	0.09	0.07	0.07
+0.4	0.06	0.08	0.10	0.17	0.09	0.05	0.08	0.12	0.09

Table 125: (n=1000) Pearson correlations of ASAM and ATE estimates across different learners. Correlations are calculated for each of the m=1000 data sets using the absolute ATE estimation error and the ASAM from LR-Stack, GBM-Stack, SL, LR, RLR, NNET, AVNNET, Twang-GBM and the two models KNN and BAG-CART, respectively. The average of the correlations in the m =1000 data sets is reported for each scenario.

	A	B	C	D	E	F	G	H	Mean
-1.2	0.34	0.40	0.38	0.32	0.36	0.35	0.34	0.31	0.35
-1.0	0.38	0.40	0.38	0.32	0.35	0.36	0.38	0.33	0.36
-0.8	0.41	0.41	0.40	0.34	0.37	0.36	0.39	0.37	0.38
-0.6	0.44	0.42	0.43	0.37	0.38	0.37	0.39	0.39	0.40
-0.4	0.47	0.42	0.44	0.39	0.40	0.40	0.42	0.41	0.42
+0.4	0.55	0.51	0.50	0.49	0.48	0.46	0.51	0.47	0.50

Table 126: (n=1000) Pearson correlations between the ASAM values and the absolute ATE estimation errors over m=1000 simulations of each model within each scenario (A-H).

	A	B	C	D	E	F	G	H	Mean
AVNNET	0.01	0.23	0.22	0.20	0.27	0.24	0.07	0.22	0.18
BAG-CART	0.29	0.32	0.24	0.20	0.32	0.21	0.25	0.27	0.26
BOOSTLR	0.25	0.19	0.22	0.23	0.26	0.21	0.24	0.30	0.24
GBM	0.03	0.05	0.08	0.06	0.06	0.05	0.05	0.04	0.05
GBM-Stack	-0.01	0.00	0.05	0.01	0.04	0.04	0.01	0.02	0.02
KNN	0.18	0.19	0.12	0.18	0.18	0.17	0.19	0.23	0.18
LR	0.02	0.00	-0.02	0.12	0.13	0.11	0.03	0.06	0.06
LR-Stack	0.01	0.01	0.09	0.09	0.21	0.13	0.04	0.02	0.07
NB	0.17	0.17	0.06	0.19	0.21	0.16	0.14	0.14	0.15
NNET	0.01	0.23	0.16	0.17	0.20	0.25	0.12	0.21	0.17
RF	-0.01	0.07	0.17	0.07	0.12	0.05	0.07	0.11	0.08
RLR	-0.01	-0.01	-0.03	0.09	0.09	0.09	0.01	0.04	0.04
SC	0.03	0.07	0.01	0.10	0.01	0.05	-0.02	0.05	0.04
SL	-0.04	0.01	-0.02	0.01	0.05	0.02	-0.01	0.03	0.01
SVM	0.03	0.02	-0.00	0.04	0.08	0.04	-0.01	0.05	0.03
Twang-GBM	-0.01	0.04	0.00	0.06	-0.00	0.03	-0.02	0.06	0.02
True PS	-0.01	0.20	0.12	0.04	0.23	0.20	0.14	0.30	0.15

Table 127: (n=3000; mse PS) Average of the mean squared errors of the simulated (“true”) propensity scores and the model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively in data setup n=3000.

	A	B	C	D	E	F	G	H	Mean
AVN	0.0022	0.0043	0.0074	0.0060	0.0059	0.0069	0.0086	0.0126	0.0067
BAG-CART	0.0229	0.0245	0.0264	0.0245	0.0258	0.0272	0.0294	0.0267	0.0259
BOOSTLR	0.0416	0.0461	0.0421	0.0511	0.0538	0.0527	0.0545	0.0616	0.0504
GBM	0.0034	0.0043	0.0066	0.0081	0.0079	0.0112	0.0126	0.0153	0.0087
GBM-Stack	0.0029	0.0049	0.0057	0.0070	0.0069	0.0085	0.0084	0.0126	0.0071
KNN	0.0386	0.0453	0.0564	0.0376	0.0409	0.0398	0.0560	0.0491	0.0455
LR	0.0009	0.0147	0.0459	0.0113	0.0147	0.0181	0.0545	0.0323	0.0241
LR-Stack	0.0030	0.0050	0.0057	0.0081	0.0068	0.0095	0.0085	0.0121	0.0073
NB	0.0100	0.0119	0.0080	0.0181	0.0183	0.0246	0.0234	0.0436	0.0198
NNET	0.0013	0.0052	0.0089	0.0073	0.0067	0.0094	0.0102	0.0157	0.0081
RF	0.0100	0.0112	0.0179	0.0115	0.0118	0.0141	0.0189	0.0189	0.0143
RLR	0.0010	0.0148	0.0460	0.0114	0.0149	0.0182	0.0546	0.0323	0.0241
SC	0.0129	0.0273	0.0544	0.0290	0.0318	0.0372	0.0636	0.0429	0.0374
SL	0.0019	0.0043	0.0062	0.0063	0.0064	0.0081	0.0087	0.0129	0.0069
SVM	0.0074	0.0077	0.0093	0.0081	0.0075	0.0094	0.0102	0.0166	0.0095
Twang-GBM	0.0041	0.0045	0.0059	0.0059	0.0058	0.0083	0.0102	0.0125	0.0072
True PS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 128: (n=3000; ks PS) Kolmogorov-Smirnov test statistics comparing the distribution of the collective simulated (“true”) propensity scores of the m=1000 simulated data sets and the collective model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively in data setup n=3000.

	A	B	C	D	E	F	G	H	Mean
AVN	0.0325	0.0173	0.0284	0.0174	0.0242	0.0158	0.0387	0.0376	0.0265
BAG-CART	0.0915	0.0853	0.0674	0.1125	0.1053	0.0770	0.0602	0.0799	0.0849
BOOSTLR	0.2956	0.3102	0.3184	0.2597	0.2333	0.2314	0.2934	0.3041	0.2808
GBM	0.0365	0.0367	0.0519	0.0784	0.0458	0.0886	0.0508	0.0907	0.0599
GBM-Stack	0.0629	0.0420	0.0546	0.1067	0.0712	0.0943	0.0596	0.1150	0.0758
KNN	0.1640	0.1469	0.1092	0.1816	0.1702	0.1425	0.0990	0.1611	0.1468
LR	0.0057	0.0537	0.2032	0.0506	0.0695	0.0488	0.2351	0.1338	0.1000
LR-Stack	0.0491	0.0568	0.0479	0.1213	0.0941	0.1105	0.0638	0.1460	0.0862
NB	0.1360	0.1199	0.0576	0.1238	0.1291	0.0433	0.0847	0.2747	0.1211
NNET	0.0185	0.0132	0.0163	0.0062	0.0102	0.0076	0.0210	0.0191	0.0140
RF	0.0964	0.0891	0.0927	0.0625	0.0853	0.0613	0.0857	0.0534	0.0783
RLR	0.0037	0.0597	0.2121	0.0562	0.0752	0.0569	0.2461	0.1427	0.1066
SC	0.2325	0.2694	0.4232	0.2886	0.2941	0.3048	0.4482	0.3454	0.3258
SL	0.0482	0.0373	0.0783	0.0722	0.0502	0.0817	0.0689	0.0901	0.0658
SVM	0.2043	0.1033	0.0518	0.1567	0.1114	0.1293	0.0496	0.1831	0.1237
Twang-GBM	0.0477	0.0407	0.0674	0.0808	0.0636	0.0830	0.0854	0.1442	0.0766
True PS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 129: (n=2000; mse PS) Average of the mean squared errors of the simulated (“true”) propensity scores and the model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively in data setup n=2000.

	A	B	C	D	E	F	G	H	Mean
AVN	0.0027	0.0065	0.0104	0.0085	0.0080	0.0102	0.0115	0.0160	0.0092
BAG-CART	0.0243	0.0261	0.0283	0.0263	0.0274	0.0294	0.0315	0.0289	0.0278
BOOSTLR	0.0425	0.0473	0.0451	0.0528	0.0546	0.0547	0.0563	0.0621	0.0519
GBM	0.0046	0.0057	0.0081	0.0099	0.0100	0.0136	0.0152	0.0181	0.0107
GBM-Stack	0.0040	0.0066	0.0070	0.0092	0.0090	0.0115	0.0111	0.0163	0.0093
KNN	0.0400	0.0480	0.0611	0.0395	0.0431	0.0423	0.0606	0.0524	0.0484
LR	0.0014	0.0152	0.0464	0.0118	0.0152	0.0186	0.0552	0.0329	0.0246
LR-Stack	0.0048	0.0078	0.0081	0.0107	0.0095	0.0156	0.0118	0.0158	0.0105
NB	0.0105	0.0121	0.0087	0.0191	0.0188	0.0255	0.0232	0.0441	0.0203
NNET	0.0020	0.0076	0.0134	0.0103	0.0095	0.0141	0.0144	0.0207	0.0115
RF	0.0104	0.0120	0.0185	0.0126	0.0130	0.0155	0.0203	0.0204	0.0153
RLR	0.0015	0.0154	0.0465	0.0120	0.0155	0.0187	0.0553	0.0331	0.0247
SC	0.0130	0.0273	0.0545	0.0290	0.0319	0.0374	0.0633	0.0431	0.0374
SL	0.0031	0.0059	0.0078	0.0085	0.0084	0.0112	0.0115	0.0161	0.0090
SVM	0.0076	0.0088	0.0119	0.0090	0.0086	0.0104	0.0131	0.0192	0.0111
Twang-GBM	0.0057	0.0061	0.0073	0.0074	0.0073	0.0098	0.0115	0.0135	0.0086
True PS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 130: (n=2000; ks PS) Kolmogorov-Smirnov test statistics comparing the distribution of the collective simulated ("true") propensity scores of the m=1000 simulated data sets and the collective model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively in data setup n=2000.

	A	B	C	D	E	F	G	H	Mean
AVN	0.0362	0.0185	0.0315	0.0160	0.0229	0.0200	0.0421	0.0407	0.0285
BAG-CART	0.0917	0.0854	0.0640	0.1095	0.1045	0.0719	0.0618	0.0768	0.0832
BOOSTLR	0.2926	0.3163	0.3109	0.2689	0.2369	0.2334	0.2887	0.3014	0.2811
GBM	0.0355	0.0362	0.0554	0.0782	0.0502	0.0929	0.0602	0.0998	0.0636
GBM-Stack	0.0662	0.0437	0.0594	0.1149	0.0708	0.0991	0.0633	0.1245	0.0802
KNN	0.1636	0.1463	0.1056	0.1799	0.1702	0.1404	0.0981	0.1584	0.1453
LR	0.0076	0.0517	0.2011	0.0499	0.0680	0.0476	0.2298	0.1336	0.0987
LR-Stack	0.0451	0.0546	0.0461	0.1065	0.0914	0.1044	0.0593	0.1391	0.0808
NB	0.1349	0.1095	0.0551	0.1291	0.1308	0.0411	0.0772	0.2710	0.1186
NNET	0.0246	0.0201	0.0218	0.0218	0.0113	0.0121	0.0199	0.0189	0.0188
RF	0.0662	0.0573	0.0632	0.0389	0.0561	0.0585	0.0880	0.0563	0.0606
RLR	0.0067	0.0604	0.2138	0.0587	0.0767	0.0584	0.2440	0.1455	0.1080
SC	0.2303	0.2664	0.4222	0.2869	0.2937	0.3036	0.4426	0.3463	0.3240
SL	0.0436	0.0298	0.1019	0.0573	0.0432	0.0725	0.0817	0.1045	0.0668
SVM	0.1841	0.0976	0.0592	0.1482	0.0996	0.1244	0.0573	0.1886	0.1199
Twang-GBM	0.0412	0.0355	0.0626	0.0727	0.0566	0.0806	0.0807	0.1377	0.0709
True PS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 131: (n=1000; mse PS) Average of the mean squared errors of the simulated ("true") propensity scores and the model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively in data setup n=1000.

	A	B	C	D	E	F	G	H	Mean
AVN	0.0059	0.0123	0.0186	0.0136	0.0134	0.0181	0.0202	0.0242	0.0158
BAG-CART	0.0272	0.0297	0.0321	0.0300	0.0312	0.0335	0.0364	0.0338	0.0317
BOOSTLR	0.0468	0.0524	0.0517	0.0564	0.0584	0.0597	0.0621	0.0657	0.0566
GBM	0.0081	0.0096	0.0125	0.0144	0.0151	0.0186	0.0212	0.0244	0.0155
GBM-Stack	0.0073	0.0118	0.0139	0.0137	0.0140	0.0178	0.0189	0.0240	0.0152
KNN	0.0430	0.0535	0.0717	0.0428	0.0476	0.0467	0.0694	0.0581	0.0541
LR	0.0029	0.0169	0.0483	0.0131	0.0166	0.0200	0.0566	0.0340	0.0260
LR-Stack	0.0088	0.0127	0.0138	0.0152	0.0155	0.0217	0.0183	0.0241	0.0163
NB	0.0130	0.0138	0.0118	0.0210	0.0202	0.0279	0.0251	0.0451	0.0222
NNET	0.0045	0.0157	0.0266	0.0143	0.0156	0.0221	0.0283	0.0328	0.0200
RF	0.0124	0.0150	0.0220	0.0156	0.0164	0.0195	0.0253	0.0244	0.0188
RLR	0.0030	0.0170	0.0483	0.0134	0.0170	0.0201	0.0565	0.0342	0.0262
SC	0.0133	0.0278	0.0550	0.0290	0.0320	0.0375	0.0635	0.0432	0.0377
SL	0.0044	0.0090	0.0137	0.0106	0.0112	0.0150	0.0188	0.0215	0.0130
SVM	0.0096	0.0116	0.0183	0.0114	0.0115	0.0138	0.0201	0.0245	0.0151
Twang-GBM	0.0101	0.0106	0.0114	0.0112	0.0116	0.0137	0.0153	0.0168	0.0126
True PS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 132: (n=1000; ks PS) Kolmogorov-Smirnov test statistics comparing the distribution of the collective simulated (“true”) propensity scores of the m=1000 simulated data sets and the collective model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively in data setup n=1000.

	A	B	C	D	E	F	G	H	Mean
AVN	0.0570	0.0201	0.0352	0.0387	0.0245	0.0258	0.0483	0.0384	0.0360
BAG-CART	0.0963	0.0878	0.0622	0.1067	0.1007	0.0685	0.0672	0.0843	0.0842
BOOSTLR	0.2822	0.3074	0.2974	0.2881	0.2598	0.2492	0.2819	0.2713	0.2797
GBM	0.0383	0.0319	0.0469	0.0627	0.0491	0.0887	0.0635	0.0909	0.0590
GBM-Stack	0.0665	0.0501	0.0640	0.1115	0.0774	0.1063	0.0725	0.1396	0.0860
KNN	0.1686	0.1487	0.1148	0.1787	0.1670	0.1424	0.0980	0.1609	0.1474
LR	0.0151	0.0491	0.1868	0.0464	0.0687	0.0442	0.2166	0.1258	0.0941
LR-Stack	0.0485	0.0491	0.0471	0.0884	0.0645	0.0984	0.0552	0.1435	0.0743
NB	0.1476	0.1052	0.0526	0.1367	0.1295	0.0385	0.0697	0.2578	0.1172
NNET	0.0416	0.0314	0.0323	0.0699	0.0379	0.0374	0.0264	0.0371	0.0393
RF	0.0336	0.0389	0.0585	0.0466	0.0552	0.0711	0.0902	0.0580	0.0565
RLR	0.0086	0.0641	0.2082	0.0614	0.0839	0.0624	0.2406	0.1465	0.1095
SC	0.2272	0.2633	0.4105	0.2840	0.2911	0.3023	0.4326	0.3410	0.3190
SL	0.0466	0.0508	0.1026	0.0694	0.0618	0.0824	0.1085	0.1261	0.0810
SVM	0.1578	0.0855	0.0851	0.1346	0.0909	0.1159	0.0825	0.1984	0.1188
Twang-GBM	0.0238	0.0196	0.0418	0.0594	0.0390	0.0742	0.0643	0.1118	0.0542
True PS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 133: (mse PS) Average of the mean squared errors of the simulated (“true”) propensity scores and the model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively and averaged over the three data size setups.

	A	B	C	D	E	F	G	H	Mean
GBM-Stack	0.0047	0.0078	0.0089	0.0100	0.0099	0.0126	0.0128	0.0177	0.0105
LR-Stack	0.0055	0.0085	0.0092	0.0113	0.0106	0.0156	0.0129	0.0173	0.0114
SL	0.0031	0.0064	0.0092	0.0085	0.0087	0.0114	0.0130	0.0169	0.0096
AVN	0.0036	0.0077	0.0121	0.0094	0.0091	0.0118	0.0134	0.0176	0.0106
NNET	0.0026	0.0095	0.0163	0.0106	0.0106	0.0152	0.0176	0.0231	0.0132
RLR	0.0018	0.0158	0.0469	0.0123	0.0158	0.0190	0.0555	0.0332	0.0250
LR	0.0018	0.0156	0.0469	0.0120	0.0155	0.0189	0.0554	0.0331	0.0249
Twang-GBM	0.0067	0.0071	0.0082	0.0082	0.0082	0.0106	0.0123	0.0143	0.0094

Table 134: (PS Kolmogorov-Smirnov) Kolmogorov-Smirnov test statistics comparing the distribution of the collective simulated (“true”) propensity scores of the m=1000 simulated data sets and the collective model estimated propensity scores over the m=1000 simulated data sets, for each scenario (A-H) respectively, averaged over the three data size setups.

	A	B	C	D	E	F	G	H	Mean
GBM-Stack	0.0652	0.0452	0.0594	0.1110	0.0732	0.0999	0.0651	0.1264	0.0807
LR-Stack	0.0476	0.0535	0.0470	0.1054	0.0834	0.1044	0.0594	0.1429	0.0804
SL	0.0461	0.0393	0.0943	0.0663	0.0517	0.0789	0.0864	0.1069	0.0712
AVN	0.0419	0.0187	0.0317	0.0240	0.0239	0.0205	0.0430	0.0389	0.0303
NNET	0.0282	0.0216	0.0235	0.0326	0.0198	0.0190	0.0224	0.0250	0.0240
RLR	0.0063	0.0614	0.2114	0.0588	0.0786	0.0592	0.2435	0.1449	0.1080
LR	0.0095	0.0515	0.1970	0.0489	0.0688	0.0469	0.2272	0.1311	0.0976
Twang-GBM	0.0375	0.0319	0.0572	0.0710	0.0531	0.0793	0.0768	0.1312	0.0672

13. Appendix B - PSY 101-Data Analysis

13.1. Psy 101 – Data description

Table 135: Variable description for the PSY 101 course data.

Variable	Label	Class		Description
Outcome Variable				
Grade	Pass	1832	(84.3%)	Pass at C- cutoff, and
	DFW	341	(15.7%)	fail DFW.
Categorical Variables				
Course number	22811 - 0	245	(11.3%)	Number of the course.
	22812 - 1	497	(22.9%)	
	22868 -2	271	(12.5%)	
	22869 - 3	485	(22.3%)	
	22942 - 4	186	(8.6%)	
	22943 - 5	489	(22.5%)	
Period	2015 - 0	742	(34.1%)	The period when the
	2016 - 1	746	(34.3%)	student took the course.
	2017 -2	675	(31.6%)	
Entry Term	2013 - 0	124	(5.7%)	Year first attendance in any
	2014 - 1	322	(14.8%)	term of the regular sessions
	2015 - 2	761	(35%)	at the reporting California
	2016 - 3	646	(30%)	State University.
	2017 - 4	320	(14.7%)	
URM	No - 0	1414	(65.1%)	Underrepresented Minority
	Yes - 1	759	(34.9%)	
Gender	Female - 0	1403	(64.6%)	Self-identified gender.
	Male - 1	770	(35.4%)	
Parent 1	NoHighSchool - 0	94	(4.3%)	Highest education of
Education	SomeHighSchool - 1	105	(4.8%)	first parent.
	HSGraduate - 2	340	(15.6%)	
	SomeCollege - 3	370	(17%)	
	2YearCollege - 4	150	(6.9%)	
	4YearCollege - 5	712	(32.8%)	
	Postgraduate - 6	340	(15.6%)	
	Unknown - 7	87	(4%)	

Students self-report their demographic information when they apply to SDSU. This includes gender, where 64.6% were identified as female and 35.4% were identified as male. The variables Parent 1 Education and Parent 2 Education indicate the highest education of both parents. We have 10% of the students that have the military indicator, which includes a verified or self-reported student that has been on active duty in the U.S. military and a student dependent of a U.S. active-duty service member. There are 32.8% of students that graduated from a school that is inside SDSU's service area. County reflects student's Institution of origin county, where

Variable	Label	Class		Description
Categorical Variables				
Parent 2 Education	NoHighSchool - 0	108	(5%)	Highest education of second parent.
	SomeHighSchool - 1	98	(4.5%)	
	HSGraduate - 2	351	(16.2%)	
	SomeCollege - 3	386	(17.8%)	
	2YearCollege - 4	151	(6.9%)	
	4YearCollege - 5	646	(29.7%)	
	Postgraduate - 6	277	(12.7%)	
	Unknown - 7	156	(7.2%)	
Military	Yes - 0	224	(1.0%)	Has had active duty or dependent.
	No - 1	1949	(99%)	
In Service Area	Yes - 0	712	(32.8%)	Graduated from a school that is inside SDSU's area.
	No - 1	1461	(67.2%)	
County Name	Los Angeles - 0	210	(10%)	Institution of Origin County.
	Orange - 1	138	(6.4%)	
	Others California - 2	605	(27.8%)	
	Out of State - 3	317	(14.6%)	
	Riverside - 4	101	(4.6%)	
	San Diego - 5	802	(36.9%)	
HS Grad Year	2013 - 0	163	(7.5%)	Year student graduated from High School.
	2014 - 1	334	(15.4%)	
	2015 - 2	746	(34.3%)	
	2016 - 3	621	(28.6%)	
	2017 - 4	309	(14.2%)	
AP Calculus	Attend Fail - 0	199	(9.2%)	Advanced placement examination Calculus.
	Attend Success - 1	236	(10.9%)	
	Non Attend - 2	1738	(79.9%)	
AP Statistics	Attend Fail - 0	129	(5.9%)	Advanced placement examination Statistics.
	Attend Success - 1	137	(6.3%)	
	Non Attend - 2	1907	(87.8%)	
AP Chemistry	Attend - 0	120	(5.5%)	Advanced placement examination Chemistry.
	Non Attend - 1	2053	(94.5%)	
AP Biology	Attend Fail - 0	103	(4.7%)	Advanced placement examination Biology.
	Attend Success - 1	138	(6.3%)	
	Non Attend - 2	1932	(89%)	

36.9% are from San Diego. Student's age has a mean of 18.11 years with a standard deviation of 1.25

The term when the student started at SDSU (Entry Term) ranges from 2013 to 2017. We merged the years from 2009 and 2012 into 2013 to prevent near-zero variance variables. The years that the students graduated from high school (HS Grad Year) go from 2013 to 2017, but again we merged the years before 2013 since they did not have many observations. The AP course variables (AP Calculus, AP Statistics, AP Chemistry, AP Biology, AP English Language, AP Literature) indicate if a student took that Advanced Placement course and then whether they

Variable	Label	Class		Description
Categorical Variables				
AP English	Attend Fail - 0	257	(11.8%)	Advanced placement
Language	Attend Success - 1	398	(18.3%)	examination English
	Non Attend - 2	1518	(69.9%)	Language.
AP English	Attend Fail - 0	250	(11.5%)	Advanced placement
Literature	Attend Success - 1	268	(12.3%)	examination English
	Non Attend - 2	1655	(76.2%)	Literature.
Fall Both	No - 0	204	(9.4%)	Proficient Math and
	Yes - 1	1969	(90.6%)	English by beginning
				of first Fall.
HS Math	No - 0	186	(8.6%)	Proficient in Math
	Yes - 1	1987	(91.4%)	by end of HS.
HS English	No - 0	138	(6.4%)	Proficient in English
	Yes - 1	2035	(93.6%)	by end of HS.
Compact	Yes - 0	273	(12.6%)	Compact Scholars
	No - 1	1900	(87.4%)	program participant.
EOP	Yes - 0	176	(9%)	Education Opportunity
	No - 1	1997	(91%)	Program.
FAST	Yes - 0	124	(5.7%)	FAST program.
	No - 1	2049	(94.3%)	participant
Res Learning	No - 0	1628	(74.9%)	Residential Learning
Community	Yes - 1	545	(25.1%)	Community program.
Term 1 SIMS	Arts & Letters - 0	90	(4.1%)	College assigned in
College	Business - 1	204	(9.4%)	term 1.
	Education - 2	102	(4.7%)	
	Engineering - 3	173	(8%)	
	Health &			
	Human Services - 4	547	(25.2%)	
	Professional Studies			
	& Fine Arts - 5	193	(8.9%)	
	Sciences - 6	573	(26.4%)	
	UG Studies - 3	291	(13.3%)	

scored a 3 or above on the AP exam for that course. About 91.4% of students were proficient in math by the end of high school (HS Math), about 93.6% were proficient in English (HS English), and about 90.6% of students were proficient in both English and Math by the beginning of the first fall (Fall Both). The mean of the incoming GPA is 3.67, with a standard deviation of 0.29. Students either took the SAT test, ACT test, or both SAT and ACT test. Therefore, the SAT variable contains the maximum of either the received scores on the SAT test or the converted score of the ACT test. The transferable units earned at colleges (Incoming Units) for transfer students has a mean of 9.24 with a standard deviation of 13.04. The number of units enrolled (EOT Term Units Enroll) has a mean of 15.06 and a standard deviation of 1.78. And the mean for the number of units earn (EOT Term Units Earn) after the respective fall semester was done is 11.85 with a standard deviation of 2.63. Units earned in PSY 101 were not included in the

Variable	Label	Class		Description
Categorical Variables				
Term1 Pre Major Status	Major - 0 Pre Major - 1	298 1875	(13.7%) (86.3%)	Student is a premajor.
Term1 Housing	Housing - 0 Non Housing - 1	1586 587	(73%) (27%)	Student resided in campus residence.
SI	No - 0 Yes - 1	1262 911	(58.1%) (41.9%)	Supplemental Instruction.
Continuous Variables				
Variable		mean	sd	Description
Age	Age Year	18.11	1.25	Age in years.
SAT Comp Conv	SAT CompConv	1138.07	134.49	Combined SAT Verbal and Math.
Incoming GPA	Incoming GPA	3.67	0.29	GPA for admission California State.
Incoming Units	Incoming Units	9.24	13.04	Transferable units earned at colleges.
EOT Term Units Enroll	EOT1 UnitsEnroll	15.06	1.78	Number of units enrolled.
EOT Term Units Earn	EOT1 UnitsEarn	11.85	2.63	Number of units earned.

We divide the data into response, categorical, and continuous variables. For the categorical variables we present the number of students and in parenthesis percentage break down of each category. For the continuous variables we provide the mean and the standard deviation.

Term Units Earn variable. The college where each student was assigned in term 1 (Term 1 SIMS College) has ten categories, where the majority of the observations are 26.4% and 25.2% for Sciences and Human Services, respectively.

The Equal Opportunity Program (EOP) is for first-generation and low-income students, mostly from underrepresented minority (URM) groups, and it includes about 9% of the students. It provides academic, financial, and counseling support with an eye on student success and retention. The URM is defined by the CSU system as students who are Black/ African-American, Latino/Hispanic, and American Indian/Native American, 34.9%. The Compact for Success scholarship program has about 12.6% of the students. It is a program created to provide students the required Math and English skills needed to succeed in college. These students start in 7th grade and continue until they graduate high school to participate in activities that will prepare them. This program also aims to improve retention of students. About 5.7% of the students are participants in the Freshmen academic success track (FAST). This program, offered in the summer prior to entering SDSU, places freshmen admitted to the university who have been placed in a Written Communication course with additional academic support. Around 25.1% students are participants of the residential learning community program (Res Learning Community).

13.2. PSY 101 – Covariate balance

In the unweighted data, there were eight covariates with standardized mean differences greater than 0.10: Entry Term (0.158), HS GradYr (0.158), County Name (0.153), Gender (0.146), Parent2 Edu (0.137), Compact (0.112), FAST (0.104) and Schnum (0.10). Further, 12 covariates had standardized mean differences between 0.10 and 0.05, and the remaining covariates lower than 0.05. The average of the standardized absolute mean differences (ASAM) over the 33 covariates in the unweighted data was 0.066. Using IPTW weights with the GBM-Stack estimated propensity scores substantially reduced the smd in the covariates. After weighting, only one covariate, Parent2 Education (0.100), had standardized mean difference greater or equal than 0.1. Nine covariates were between 0.1 and 0.05, leading to an overall ASAM (including all 33 covariates) of 0.040. Table B.136 and Table B.140 present the standardized differences of all covariates between the treatment SI attendance group and the control group before and after weighting.

None of the six continuous covariates had unweighted standardized mean difference greater than 0.1, and also the quadratic and cubic terms were lower than 0.1 before weighting, with an average of 0.034 and 0.043, respectively. Weighting slightly reduced the averaged standardized mean difference of the quadratic and cubic terms of the continuous covariates to an average of 0.031 and 0.037, respectively. All values are compared in Table B.141. Unweighted, the mean Kolmogorov-Smirnov statistics (K-S) across all covariates was 0.0154, with a maximum at 0.0696, which could both be decreased by weighting. After weighting, a mean K-S of 0.0104 was obtained, averaged over all covariates. The maximum K-S was 0.0483 for the combined SAT score covariate. The detailed K-S statistics for all covariates are given in Table B.143, with a summary in Table B.142.

The weight distributions (computed through Formula 3, Section 2.1) were well centered between 1 and 2, with maxima weights below 10 in both groups. The weights in the treatment group were slightly higher compared to the control group (Table B.146). The overall relatively low weights show that no observations overly impact the treatment effect estimation due to large weights.

The comparisons show that there was moderate imbalance in the data before weighting, which could be diminished by IPTW, using our GBM-Stack estimated propensity scores. For the purpose of our study, we are confident about the balance in the data after weighting. Franklin et al. (2014) propose a post-matching or post-weighting C-statistic as a balance score which could be additionally evaluated if the balance of the data set is not as clear.

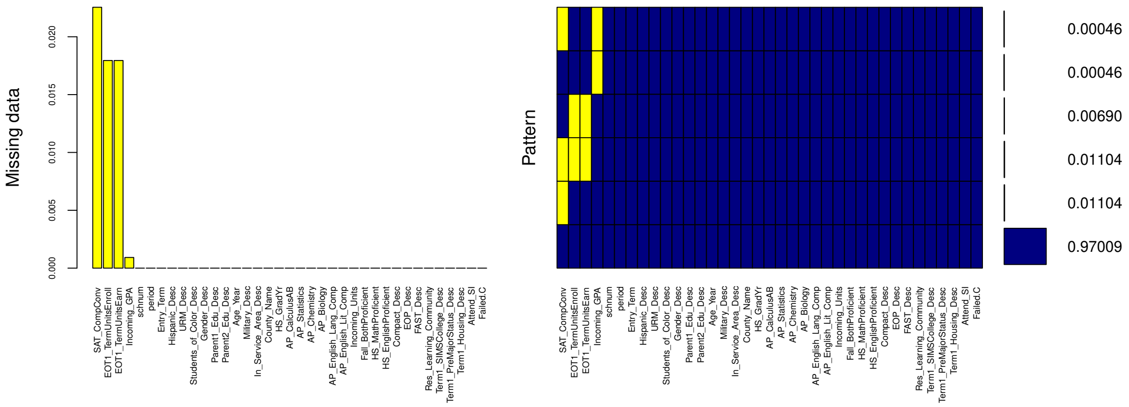

Figure 7: Pattern of missing vales in the PSY101 data set.

Table 136: Standardized mean differences of the covariates, stratified by the treatment variable SI attendance. For categorical variables quantity and (proportion in %) is listed for each label in each group. For binary variables we just indicate one label. For numerical variables the stratified mean and (sd) is listed.

	NO	YES	SMD
Count	1262	911	
Schnum (%)			0.100
22811	157 (12.4)	88 (9.7)	
22812	286 (22.7)	211 (23.2)	
22868	154 (12.2)	117 (12.8)	
22869	283 (22.4)	202 (22.2)	
22942	110 (8.7)	76 (8.3)	
22943	272 (21.6)	217 (23.8)	
period (%)			0.052
20154	443 (35.1)	299 (32.8)	
20164	437 (34.6)	319 (35.0)	
20174	382 (30.3)	293 (32.2)	
Entry Term (%)			0.158
20134	74 (5.9)	50 (5.5)	
20144	194 (15.4)	128 (14.1)	
20154	464 (36.8)	297 (32.6)	
20164	372 (29.5)	274 (30.1)	
20174	158 (12.5)	162 (17.8)	
URM Desc = Underrepresented (%)	443 (35.1)	316 (34.7)	0.009
Gender Desc = Male (%)	484 (38.4)	286 (31.4)	0.146
Parent1 Edu Desc (%)			0.089
2-Year College Grad (5)	83 (6.6)	67 (7.4)	
4-Year College Grad (6)	412 (32.6)	300 (32.9)	
HS Graduate (3)	206 (16.3)	134 (14.7)	
No High School (1)	55 (4.4)	39 (4.3)	
Others Unknown	50 (4.0)	37 (4.1)	
Postgraduate (7)	185 (14.7)	130 (14.3)	
Some College (4)	204 (16.2)	166 (18.2)	
Some High School (2)	67 (5.3)	38 (4.2)	
Parent2 Edu Desc (%)			0.137
2-Year College Grad (5)	88 (7.0)	63 (6.9)	
4-Year College Grad (6)	368 (29.2)	278 (30.5)	
HS Graduate (3)	207 (16.4)	144 (15.8)	
No High School (1)	72 (5.7)	36 (4.0)	
Others Unknown	97 (7.7)	59 (6.5)	
Postgraduate (7)	148 (11.7)	129 (14.2)	
Some College (4)	218 (17.3)	168 (18.4)	
Some High School (2)	64 (5.1)	34 (3.7)	

Table 138: Standardized mean differences of the covariates, stratified by the treatment variable SI attendance. For categorical variables quantity and (proportion in %) is listed for each label in each strata. For binary variables we just indicate one label. For numerical variables the stratified mean and (sd) is listed.

Age Year (mean (sd))	18.10 (1.12)	18.11 (1.41)	0.008
Military Desc = Not Military (%)	1128 (89.4)	821 (90.1)	0.024
In Service Area = Non-Local (%)	833 (66.0)	628 (68.9)	0.063
County Name (%)			0.153
LosAngeles	122 (9.7)	88 (9.7)	
Orange	62 (4.9)	76 (8.3)	
Others California	348 (27.6)	257 (28.2)	
Out of State	181 (14.3)	136 (14.9)	
Riverside	61 (4.8)	40 (4.4)	
SanDiego	488 (38.7)	314 (34.5)	
HS GradYr (%)			0.158
2013	93 (7.4)	70 (7.7)	
2014	202 (16.0)	132 (14.5)	
2015	457 (36.2)	289 (31.7)	
2016	357 (28.3)	264 (29.0)	
2017	153 (12.1)	156 (17.1)	
SAT CompConv (mean (sd))	1142.69 (134.61)	1131.68 (134.14)	0.082
AP CalculusAB (%)			0.067
attend fail	114 (9.0)	85 (9.3)	
attend success	148 (11.7)	88 (9.7)	
Non attend	1000 (79.2)	738 (81.0)	
AP Statistics (%)			0.075
attend fail	70 (5.5)	59 (6.5)	
attend success	88 (7.0)	49 (5.4)	
Non attend	1104 (87.5)	803 (88.1)	
AP Chemistry = Non attend (%)	1193 (94.5)	860 (94.4)	0.006
AP Biology (%)			0.078
attend fail	60 (4.8)	43 (4.7)	
attend success	90 (7.1)	48 (5.3)	
Non attend	1112 (88.1)	820 (90.0)	
AP English Lang Comp (%)			0.025
attend fail	146 (11.6)	111 (12.2)	
attend success	235 (18.6)	163 (17.9)	
Non attend	881 (69.8)	637 (69.9)	
AP English Lit Comp (%)			0.038
attend fail	139 (11.0)	111 (12.2)	
attend success	155 (12.3)	113 (12.4)	
Non attend	968 (76.7)	687 (75.4)	

Table 139: Standardized mean differences of the covariates, stratified by the treatment variable SI attendance. For categorical variables quantity and (proportion in %) is listed for each label in each strata. For binary variables we just indicate one label. For numerical variables the stratified mean and (sd) is listed.

Incoming GPA (mean (sd))	3.66 (0.29)	3.68 (0.29)	0.069
Incoming Units (mean (sd))	9.42 (12.58)	8.99 (13.65)	0.033
Fall BothProficient = 1 (%)	1147 (90.9)	822 (90.2)	0.022
HS MathProficient = 1 (%)	1164 (92.2)	823 (90.3)	0.067
HS EnglishProficient = 1 (%)	1191 (94.4)	844 (92.6)	0.070
Compact = Not Compact Scholar (%)	1084 (85.9)	816 (89.6)	0.112
EOP Desc = Not EOP (%)	1158 (91.8)	839 (92.1)	0.012
FAST Desc = Not FAST (%)	1203 (95.3)	846 (92.9)	0.104
Res Learning Community = Res Learning (%)	311 (24.6)	234 (25.7)	0.024
Term1 SIMSCollege Desc (%)			0.079
Arts and Letters	53 (4.2)	37 (4.1)	
Business	128 (10.1)	76 (8.3)	
Education	54 (4.3)	48 (5.3)	
Engineering	102 (8.1)	71 (7.8)	
Health and Human Services	313 (24.8)	234 (25.7)	
Professional Studies and Fine Arts	112 (8.9)	81 (8.9)	
Sciences	334 (26.5)	239 (26.2)	
UG Studies	166 (13.2)	125 (13.7)	
Term1 PreMajorStatus Desc = Pre Major (%)	1093 (86.6)	782 (85.8)	0.022
Term1 Housing Desc = Non-Housing (%)	360 (28.5)	227 (24.9)	0.082
EOT1 TermUnitsEnroll (mean (sd))	15.07 (1.76)	15.06 (1.81)	0.001
EOT1 TermUnitsEarn (mean (sd))	11.82 (2.77)	11.88 (2.43)	0.023

Table 140: Standardized mean differences between covariates in the PSY 101 data stratified by the treatment variable SI attendance. Unweighted smd and the smd resulting from IPTW with GBM-Stack estimated propensity scores are presented. Results are ranked on the smd in the unweighted data set.

Covariate	Unweighted smd	GBM-Stack
Entry Term	0.158	0.041
HS Grad Yr	0.158	0.054
County Name	0.153	0.096
Gender Desc	0.146	0.049
Parent2 Edu Desc	0.137	0.100
Compact Desc	0.112	0.065
FAST Desc	0.104	0.047
schnum	0.100	0.091
Parent1 Edu Desc	0.089	0.077
SAT CompConv	0.082	0.083
Term1 Housing Desc	0.082	0.048
Term1 SIMSCollege Desc	0.079	0.039
AP Biology	0.078	0.052
AP Statistics	0.075	0.066
HS EnglishProficient	0.070	0.042
Incoming GPA	0.069	0.020
HS MathProficient	0.067	0.030
AP CalculusAB	0.067	0.047
In Service Area Desc	0.063	0.032
period	0.052	0.062
AP English Lit Comp	0.038	0.024
Incoming Units	0.033	0.014
AP English Lang Comp	0.025	0.018
Military Desc	0.024	0.009
Res Learning Community	0.024	0.022
EOT1 TermUnitsEarn	0.023	0.029
Fall BothProficient	0.022	0.002
Term1 PreMajorStatus Desc	0.022	0.027
EOP Desc	0.012	0.018
URM Desc	0.009	0.003
Age Year	0.008	0.019
AP Chemistry	0.006	0.009
EOT1 TermUnitsEnroll	0.001	0.011
Average	0.066	0.040

Table 141: Standardized mean differences between the squared terms of continuous covariates (smdQ2) and the cubic terms (smdQ3). Results are ranked on the smd of quadratic terms of the unweighted continous covariates.

Variable	Unw. (smdQ2)	GBM-Stack (Q2)	Unw. (smdQ3)	GBM-Stack (Q3)
SAT Q2	0.083	0.086	0.083	0.089
Inc GPA Q2	0.072	0.022	0.075	0.023
Inc Units Q2	0.021	0.027	0.040	0.040
Age Q2	0.017	0.030	0.027	0.040
EOT Earn Q2	0.006	0.004	0.026	0.013
EOT Enroll Q2	0.003	0.015	0.006	0.019
Average	0.034	0.031	0.043	0.037

Table 142: Summary of the Kolmogorov-Smirnoff test statistics to measure the difference of the empirical CDF's of the treatment and control group in the PSY101 data. GBM-Stack represents the ks-statistic in the weighted data using GBM-Stack estimated propensity scores.

Variable	Unweighted	GBM-Stack
mean K-S	0.0155	0.0104
max K-S	0.0696	0.0483

Table 143: Kolmogorov-Smirnoff test statistics listed for each continuous variable and each category for categorical variables to measure the difference of the empirical CDF's of the treatment and control group in the PSY101 data.

	Unweighted	GBM Stack
schnum 22811	0.028	0.015
schnum 22812	0.005	0.014
schnum 22868	0.006	0.017
schnum 22869	0.003	0.009
schnum 22942	0.004	0.014
schnum 22943	0.023	0.011
period 20154	0.023	0.001
period 20164	0.004	0.026
period 20174	0.019	0.024
Entry Term:20134	0.004	0.001
Entry Term:20144	0.013	0.002
Entry Term:20154	0.042	0.012
Entry Term:20164	0.006	0.019
Entry Term:20174	0.053	0.003
URM	0.004	0.001
Gender	0.070	0.023
Parent1 Edu: 2-Year College Grad (5)	0.008	0.009
Parent1 Edu: 4-Year College Grad (6)	0.003	0.008
Parent1 Edu: HS Graduate (3)	0.016	0.013
Parent1 Edu: No High School (1)	0.001	0.001
Parent1 Edu: Others Unknown	0.001	0.001
Parent1 Edu: Postgraduate (7)	0.004	0.011
Parent1 Edu: Some College (4)	0.021	0.016
Parent1 Edu: Some High School (2)	0.011	0.007
Parent2 Edu: 2-Year College Grad (5)	0.001	0.000
Parent2 Edu: 4-Year College Grad (6)	0.014	0.017
Parent2 Edu: HS Graduate (3)	0.006	0.005
Parent2 Edu: No High School (1)	0.018	0.014
Parent2 Edu: Others Unknown	0.012	0.009
Parent2 Edu: Postgraduate (7)	0.024	0.009
Parent2 Edu: Some College (4)	0.012	0.012
Parent2 Edu: Some High School (2)	0.013	0.010
Age Year	0.014	0.015
Military	0.007	0.003
In Service Area	0.029	0.015

Table 144: Follow-up: Kolmogorov-Smirnoff test statistics listed for each continuous variable and each category for categorical variables to measure the difference of the empirical CDF's of the treatment and control group in the PSY101 data.

	Unweighted	GBM Stack
County Name: LosAngeles	0.000	0.002
County Name: Orange	0.034	0.021
County Name: Others California	0.006	0.008
County Name: Out of State	0.006	0.002
County Name: Riverside	0.004	0.001
County Name: SanDiego	0.042	0.027
HS GradYr: 2013	0.003	0.005
HS GradYr: 2014	0.015	0.002
HS GradYr: 2015	0.045	0.016
HS GradYr: 2016	0.007	0.020
HS GradYr: 2017	0.050	0.008
SAT CompConv	0.039	0.048
AP CalculusAB attend fail	0.003	0.007
AP CalculusAB attend success	0.021	0.013
AP CalculusAB Non attend	0.018	0.007
AP Statistics attend fail	0.009	0.012
AP Statistics attend success	0.016	0.011
AP Statistics Non attend	0.007	0.001
AP Chemistry	0.001	0.002
AP Biology attend fail	0.000	0.000
AP Biology attend success	0.019	0.013
AP Biology Non attend	0.019	0.012
AP English Lang Comp attend fail	0.006	0.004
AP English Lang Comp attend success	0.007	0.005
AP English Lang Comp Non attend	0.001	0.001
AP English Lit Comp attend fail	0.012	0.007
AP English Lit Comp attend success	0.001	0.001
AP English Lit Comp Non attend	0.013	0.009
Incoming GPA	0.068	0.042
Incoming Units	0.043	0.030
Fall BothProficient	0.007	0.000
HS MathProficient	0.019	0.008
HS EnglishProficient	0.017	0.010
Compact	0.037	0.021
EOP	0.003	0.005
FAST	0.025	0.011

Table 145: Follow-up: Kolmogorov-Smirnoff test statistics listed for each continuous variable and each category for categorical variables to measure the difference of the empirical CDF's of the treatment and control group in the PSY101 data.

	Unweighted	GBM Stack
Res Learning Community	0.010	0.010
Term1 SIMSCollege: Arts & Letters	0.001	0.000
Term1 SIMSCollege: Business	0.018	0.006
Term1 SIMSCollege: Education	0.010	0.002
Term1 SIMSCollege: Engineering	0.003	0.001
Term1 SIMSCollege: Health & Human Services	0.009	0.006
Term1 SIMSCollege: Professional Studies & Fine Arts	0.000	0.004
Term1 SIMSCollege: Sciences	0.002	0.012
Term1 SIMSCollege: UG Studies	0.006	0.006
Term1 PreMajorStatus	0.008	0.009
Term1 Housing	0.036	0.021
EOT1 TermUnitsEnroll	0.019	0.015
EOT1 TermUnitsEarn	0.030	0.027

Table 146: Summary of the computed IPTW ATE weights based on GBM-Stack estimated propensity scores in the PSY 101 data grouped by SI attendance.

Attend SI	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
NO	1.16	1.55	1.73	1.80	1.93	4.40
YES	1.27	1.95	2.29	2.40	2.69	5.12