

RESEARCH

Open Access



Prognostic score-based model averaging approach for propensity score estimation

Daijiro Kabata^{1,2,3*}, Elizabeth A. Stuart⁴ and Ayumi Shintani²

Abstract

Background Propensity scores (PS) are typically evaluated using balance metrics that focus on covariate balance, often without considering their predictive power for the outcome. This approach may not always result in optimal bias reduction in the treatment effect estimate. To address this issue, evaluating covariate balance through prognostic scores, which account for the relationship between covariates and the outcome, has been proposed. Similarly, using a typical model averaging approach for PS estimation that minimizes prediction error for treatment status and covariate imbalance does not necessarily optimize PS-based confounding adjustment. As an alternative approach, using the averaged PS model that minimizes inter-group differences in the prognostic score may further reduce bias in the treatment effect estimate. Moreover, since the prognostic score is also an estimated quantity, model averaging in the prognostic scores can help identify a better prognostic score model. Utilizing the model-averaged prognostic scores as the balance metric for constructing the averaged PS model can contribute to further decreasing bias in treatment effect estimates. This paper demonstrates the effectiveness of the PS model averaging approach based on prognostic score balance and proposes a method that uses the model-averaged prognostic score as a balance metric, evaluating its performance through simulations and empirical analysis.

Methods We conduct a series of simulations alongside an analysis of empirical observational data to compare the performances of weighted treatment effect estimates using the proposed and existing approaches. In our examination, we separately provide four candidate estimates for the PS and prognostic score models using traditional regression and machine learning methods. The model averaging of PS based on these candidate estimators is performed to either maximize the prediction accuracy of the treatment or to minimize intergroup differences in covariate distributions or prognostic scores. We also utilize not only the prognostic scores from each candidate model but also an averaged score that best predicted the outcome, for the balance assessment.

Results The simulation and empirical data analysis reveal that our proposed model-averaging approaches for PS estimation consistently yield lower bias and less variability in treatment effect estimates across various scenarios compared to existing methods. Specifically, using the optimally averaged prognostic scores as a balance metric significantly improves the robustness of the weighted treatment effect estimates.

Discussion The prognostic score-based model averaging approach for estimating PS can outperform existing model averaging methods. In particular, the estimator using the model averaging prognostic score as a balance metric can produce more robust estimates. Since our results are obtained under relatively simple conditions, applying them to real data analysis requires adjustments to obtain accurate estimates according to the complexity and dimensionality of the data.

*Correspondence:

Daijiro Kabata
daijiro.kabata@port.kobe-u.ac.jp

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Conclusions Using the prognostic score as the balance metric for the PS model averaging enhances the performance of the treatment effect estimator, which can be recommended for a wide variety of situations. When applying the proposed method to real-world data, it is important to use it in conjunction with techniques that mitigate issues arising from the complexity and high dimensionality of the data.

Keywords Causal inference, Propensity score, Prognostic score, Model averaging, Machine learning

In observational studies, the lack of random treatment assignment leads to imbalanced covariate distributions between treatment groups, resulting in confounding that biases treatment effect estimates. To reduce the effect of the confounding, using propensity scores (PS) is one of the viable approaches [1]. PS represents the likelihood of each subject receiving the treatment, calculated from observed covariates. By matching or weighting subjects based on PS, the covariate distributions tend to become similar, thereby reducing bias in treatment effect estimates.

For PS estimation, logistic regression models are typically used, based on observed covariates. To better understand complex covariate-treatment assignment relationships, machine learning methods are increasingly employed [2, 3]. To further mitigate the risk of PS model misspecification, a model averaging approach has been suggested, which involves generating multiple PS estimates and integrating them. For instance, Pirracchio et al. proposed the *SuperLearner* approach, aggregating various machine learning-derived PS candidates based on predictive accuracy [4]. However, it's important to note that a PS model's accuracy in predicting treatment assignment doesn't always correspond to the reduction of bias in treatment effect estimates [5]. PS-based estimators aim to create a pseudo population where confounders-factors linked to both treatment assignment and outcome are effectively balanced between treatment groups. To achieve this with multiple PS models, Xie et al. recommended a model averaging approach for PS, focusing on minimizing covariate distribution imbalances between treatment groups [6]. This technique averages PS models to align covariate distributions more closely after weighting by the model-averaged PS.

Xie et al.'s approach employs balance metrics like absolute standardized mean difference (ASMD) and Kolmogorov-Smirnov (KS) statistics of observed covariates. However, Stuart et al. demonstrate that the imbalance in prognostic scores is more closely associated with bias in treatment effect estimates and recommend it as a covariate balance metric [7]. The prognostic score, indicating the baseline risk of untreated subjects, is derived from covariates that are closely related to the outcome [8, 9]. Prioritizing covariates related to the outcome is crucial for reducing bias, as opposed to those related

to treatment assignment [10–14]. Incorporating Stuart et al.'s recommendation, this paper presents a model averaging approach for PS estimation that uses the prognostic score as the balance metric. Our method aligns with Xie et al.'s model averaging approach but with a focus on the prognostic score to achieve a more efficient balance for the treatment effect estimation. Furthermore, to enhance the robustness of the proposed estimator, we introduce using the model-averaged prognostic score as the balance metric as well as using each prognostic score candidate.

In this paper, we evaluate our approach with comprehensive simulation studies and by applying it to a real-world clinical dataset.

Estimators

Propensity scores and target parameter

We consider a set of subjects' characteristics, represented by a p -dimensional vector $X_i = (X_{1i}, X_{2i}, \dots, X_{pi})$. The treatment indicator A_i specifies whether a patient received treatment (1) or not (0). The probability that each subject receives the treatment is denoted as $e(X_i)$, namely, PS. Based on the subjects' characteristics and the treatment, the outcome Y_i is modeled as $Y_i = g(A_i, X_i) + \epsilon_i$, where ϵ_i is an error term. Then, our primary target parameter, the average treatment effect on the treated subjects (ATT) is given by $\theta = E[g(1, X_i) - g(0, X_i) | A_i = 1]$.

Inverse probability weighting estimator

To estimate ATT, we can employ the inverse probability weighting (IPW) estimator, which is given by

$$\hat{\theta} = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1 - A_i) Y_i \hat{W}_i}{\sum_{i=1}^n (1 - A_i) \hat{W}_i}$$

where $\hat{W}_i$ is the weight for i -th subject given as $\hat{W}_i = \hat{e}(X_i) / (1 - \hat{e}(X_i))$ [15, 16].

Model averaging PS based on covariate balance

Xie et al. proposed a model-averaging approach that integrates multiple PS estimates to improve robustness against model misspecification in PS estimation.

Although their demonstration featured only two PS model candidates, this method can be generally applied by

$$\hat{e}(X_i) = \sum_{m=1}^M \lambda_m \hat{e}_m(X_i).$$

Here, M represents the total number of candidate models for PS, $\hat{e}_m(X_i)$ denotes PS estimate using the m -th regression model or machine learning method, and λ_m

$$wASMD^{Prog} = \left| \frac{\sum_{i=1}^n A_i \hat{g}(0, X_i)}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1 - A_i) \hat{g}(0, X_i) \hat{W}_i}{\sum_{i=1}^n (1 - A_i) \hat{W}_i} \right| / sd(\hat{g}(0, X) | A = 1).$$

is the mixing parameter. These mixing parameters adhere to the conditions $\lambda_m \in [0, 1]$ and $\sum_{m=1}^M \lambda_m = 1$. Xie et al. suggested selecting an optimal mix of these parameters to minimize covariate distribution imbalances between treatment groups after weighting with a model averaging propensity score. For balance metrics, they used the mean of the weighted ASMD and the weighted KS statistic. The weighted ASMD for each covariate X_j ($j = 1, 2, \dots, p$) is calculated as

$$wASMD^{(j)} = \left| \frac{\sum_{i=1}^n A_i X_{ji}}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1 - A_i) X_{ji} \hat{W}_i}{\sum_{i=1}^n (1 - A_i) \hat{W}_i} \right| / sd(X_j | A = 1),$$

which is defined as the absolute difference in means of a covariate across the treatment groups, divided by the standard deviation in the treated group $sd(X_j | A = 1)$ [7, 16]. In a similar manner to the weighted ASMD, the weighted KS statistic for each covariate X_j is expressed by

$$wKS^{(j)} = \sup_{x_j} |F_{n_1}(x_j | A = 1) - F_{n_0}^w(x_j | A = 0)|,$$

where $F_{n_a}(x_j | A = a) = \frac{1}{n_a} \sum_{i:A_i=a} I(X_{ij} \leq x_j)$ is the cumulative distribution function of the covariate X_j in the group with treatment $A = a$, and $F_{n_a}^w(x_j | A = a) = \frac{1}{\sum_{i:A_i=a} \hat{W}_i} \sum_{i:A_i=a} I(X_{ij} \leq x_j) \hat{W}_i$ is the weighted one. n_1 and n_0 denoting the numbers of subjects in the treated and untreated groups, respectively. $I(X_{ij} \leq x_j)$ is an indicator function that equals 1 if $X_{ij} \leq x_j$ and 0 otherwise.

The optimal set of the mixing parameter λ_m minimizing the weighted balance metric is found by the grid search method. In the grid search procedure, we provide the model averaging PSs using possible sets for λ_m and evaluate the weighted balance metrics. Then, we identify the optimal set of λ_m that results in the optimal balance.

Model averaging PS based on prognostic score balance

Based on the approach proposed by Xie et al, we propose a model averaging method using prognostic scores as the balance metric for selecting the optimal mixing parameter. The prognostic score for each subject can be defined as $g(0, X_i)$, which represents the predicted outcome for a subject who receives reference treatment. Our method selects the mixing parameter that minimizes the mean weighted ASMD of these prognostic scores, which is given by

The $\hat{g}(0, X)$ is the estimated model of the prognostic score $g(0, X)$, derived from the observed data on subjects who did not receive treatment of interest ($A = 0$). By applying the estimated model to observed covariates of each subject X_i , we can obtain an estimated prognostic score for each subject $\hat{g}(0, X_i)$.

Model averaging for prognostic score estimation

Improving the prediction accuracy of the prognostic score is generally expected to enhance the performance of estimates using them because prognostic scores are functions of covariates related to the outcome variable [10]. Thus, constructing prognostic scores that accurately predict outcomes is vital for optimal treatment effect estimates. We therefore introduce a model averaging method for prognostic score estimation, focused on outcome variable prediction accuracy. The model averaging prognostic score is defined by

$$\hat{\bar{g}}(0, X_i) = \sum_{m=1}^M \gamma_m \hat{g}_m(0, X_i),$$

where γ_m is a mixing parameter for m -th candidates for prognostic score model $\hat{g}_m(0, X_i)$. The mixing parameter for the prognostic score averaging is also determined by the grid search method, aiming to minimize the mean squared error (MSE) of the prognostic score estimates in subjects not receiving the treatment.

The IPW estimator and the weighted balance metrics for model averaging described above need to be adjusted for the estimation of the average treatment effect among the entire cohort, namely ATE, including modifications to the weighting scheme. The details of these modifications and corresponding formulas are provided in the Supplement.

Simulation experiment

To assess the efficacy of our approach and compare with the existing approaches, we conduct a series of simulation experiments. These experiments are designed to evaluate the performance of the IPW estimator using our proposed model averaging PS method based on the prognostic score. For a comprehensive comparison, we also examine the performance of the IPW estimator using a traditional PS approach without model averaging (referred to as single PS), as well as an existing model averaging PS method that emphasizes covariate balance and the predictive accuracy of treatment assignments.

Settings of simulation experiment

The simulation design is inspired by the work of Setoguchi et al., Lee et al., and Stuart et al. [2, 3, 7]. We conduct simulations for sample sizes n of 1000. First of all, we generate 10 covariates. Let $X_{j \in \{1,3,5,6,8,9\}}$ be binary variables following a Bernoulli distribution with a probability of 0.5 and $X_{j \in \{2,4,7,10\}}$ be continuous variables following a standard normal distribution. Correlations are introduced as $\text{corr}(X_1, X_5) = \text{corr}(X_3, X_8) = 0.2$ and $\text{corr}(X_2, X_6) = \text{corr}(X_4, X_9) = 0.9$. Covariates $X_{j \in \{1,2,3,4\}}$ are confounders associated with both the treatment A_i and the outcome Y_i , $X_{j \in \{5,6,7\}}$ are only associated with the treatment A_i , and $X_{j \in \{8,9,10\}}$ are only associated with the outcome Y_i . The average treatment effect θ is set as -0.4 . The relations between the variables are illustrated in Fig. 1.

The treatment indicator for each subject A_i is generated based on a Bernoulli distribution with the success probability determined by the true PS model $e(X_i)$. Similarly, the continuous outcome Y_i is determined by its true outcome model. We provide four different true models for both PS and the outcome with the following properties:

- A : additivity and linearity (main effects only);
- B : nonlinearity (3 quadratic terms);
- C : nonadditivity (4 two-way cross-product terms);

D : nonlinearity and nonadditivity (3 quadratic and 4 two-way cross-product terms).

The formulas are available in the Supplement. We examine 16 scenarios combining these four types of PS models and four types of outcome models.

Propensity and prognostic score estimators

For PS and prognostic score estimation in our simulation, we choose four candidate models encompassing both traditional regression and machine learning methods. These include logistic regression only with linear terms (LGR), classification and regression trees (CART), random forest (RF), and generalized boosted models (GBM). This selection of estimating methods aligns with those commonly used in prior research assessing the efficacy of machine learning techniques for the nuisance function estimation [2, 3, 6]. Our decision to focus on these four methods is driven by the aim to optimize computational efficiency while maintaining a robust and representative range of estimation techniques.

Each candidate model is independently estimated using the *SuperLearner* function of R specifying only each machine learning method as the learner argument [17]. Based on the individual model estimates, we compute the weight for the ATT estimation with the following three different approaches: a single PS approach utilizing each PS candidate individually, the existing model averaging PS approach optimizing covariate balance and prediction accuracy of treatment assignments, and our proposed model averaging PS approach that optimizes the balance of prognostic scores. For constructing the existing model averaging PS, we employ the mean weighted ASMD and the mean weighted KS statistic as the balance metrics similar to Xie et al. Furthermore, to examine the performance of the traditional model averaging PS based on the predictive accuracy of treatment

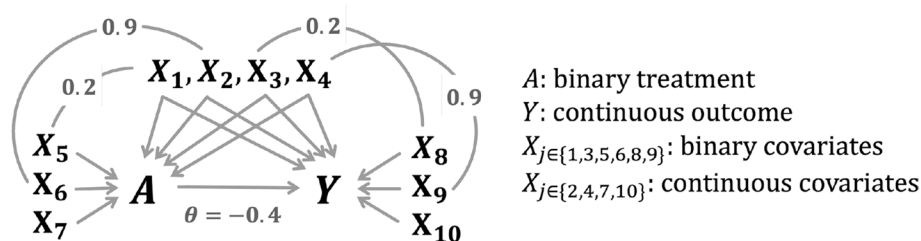


Fig. 1 Simulation Data Structure. Covariates $X_{j \in \{1,2,3,4\}}$ are confounders associated with both the treatment A and the outcome Y , $X_{j \in \{5,6,7\}}$ are instruments associated only with the treatment A , and $X_{j \in \{8,9,10\}}$ are outcome predictors associated only with the outcome Y . X_1 and X_5 , X_3 and X_8 are weakly correlated with a correlation coefficient of 0.2. X_2 and X_6 , X_4 and X_9 are strongly correlated with a correlation coefficient of 0.9. The average treatment effect θ is set at -0.4

assignment, we utilize Log Loss function, expressed as $-\frac{1}{n} \sum_{i=1}^n \{a_i \cdot \log(\hat{e}_i(X)) + (1 - a_i) \cdot \log(1 - \hat{e}_i(X))\}$, as the balance metric. This function measures the predictive accuracy and is crucial in evaluating how well the PS model predicts treatment assignment. To find out the optimal mixing parameter for the construction of the averaging PS models and the averaging prognostic score models, the grid search is conducted at an increment of 0.05. Supplemental Table 1 summarizes the series of the employed analytical approaches.

Performance metrics

We replicate the simulation procedure $R = 1,000$ times, and then, evaluate the performances of the estimators. We employ a comprehensive set of performance metrics as follows: (1) absolute bias $|\hat{\theta} - \theta|$ where $\hat{\theta}$ is the mean of estimates across all iterations, (2) empirical standard error $\sqrt{R^{-1} \sum_{r=1}^R (\hat{\theta}_r - \bar{\hat{\theta}})^2}$ where $\hat{\theta}_r$ is an estimate in each iteration r , (3) root mean squared error (RMSE) $\sqrt{R^{-1} \sum_{r=1}^R (\hat{\theta}_r - \theta)^2}$, (4) 95% confidence interval coverage rate $R^{-1} \sum_{r=1}^R I_r(\hat{\theta}_r, \hat{se}_r)$ where $I_r(\hat{\theta}_r, \hat{se}_r)$ is an indicator which takes 1 if 95% confidence interval in each iteration $\hat{\theta}_r \pm 1.96\hat{se}_r$ covers the true effect θ and 0 otherwise. The estimated standard error $\hat{se}_r$ is obtained using the sandwich method in *survey* package in R. Furthermore, we illustrate the distributions of the mixing parameters and the balance metrics.

Results

We assess the effectiveness of the model averaging approach based on the prognostic score via comparison with the single PS approach and the existing model averaging approaches based on the covariate balance and the prediction accuracy in 16 cases, each being a permutation of four true PS models and four different outcome models. To facilitate interpretation, we excerpt four representative cases that characterize each estimator's features within these 16 scenarios. Case 1 (PS A & Outcome A) features both the true PS and outcome models as simple; Case 2 (PS A & Outcome D) presents a complex true outcome model and a simple true PS model; Case 3 (PS D & Outcome A) involves a complex true PS model and a simple true outcome model; Case 4 (PS D & Outcome D) has both true PS and outcome models as complex. The performance metrics from our simulation experiment for these four representative cases are shown in Supplemental Table 2. The distributions of the estimates are illustrated in Fig. 2 and the 95% CI coverage rates are shown in Fig. 3. Furthermore, Fig. 4 illustrates the distributions of the mixing parameter λ_m for PS model averaging. Supplemental Figure 1-4 display the distributions of balance metrics for detecting λ_m . Similarly, Fig. 5 illustrates the distributions of the mixing parameter γ_m for the prognostic score averaging. Supplemental Figure 5 presents the distributions of the MSE of the estimated prognostic scores for identifying γ_m on the prognostic score model averaging approach.

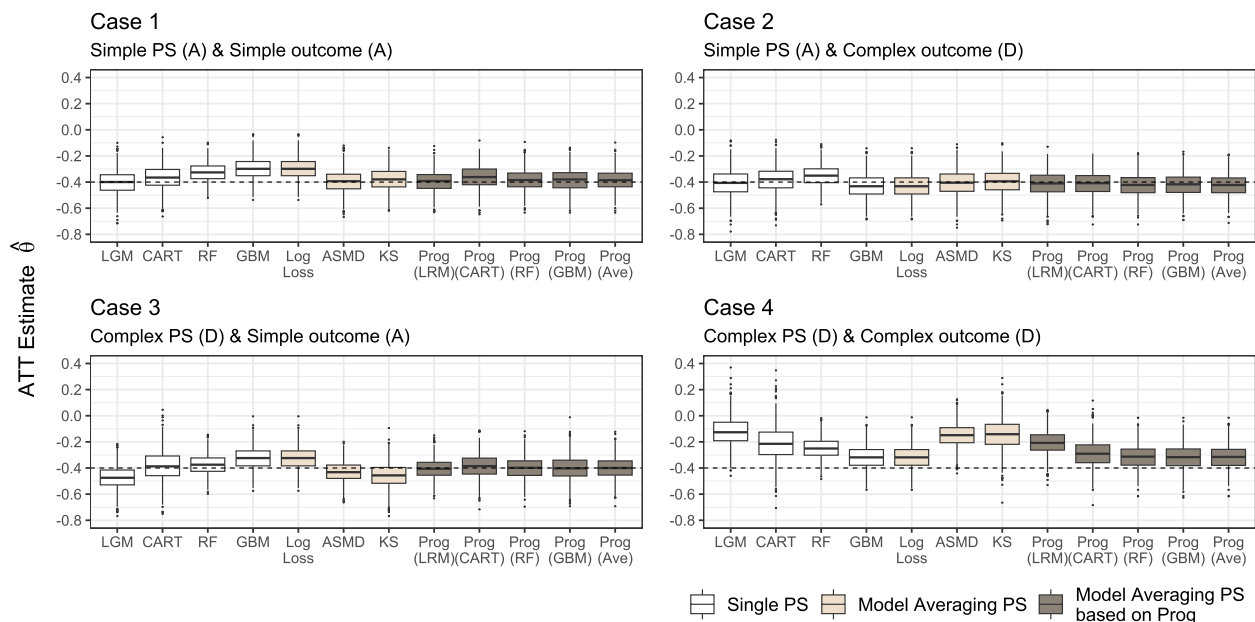


Fig. 2 Distributions of Estimates. Abbreviations: ATT, average treatment effect on treated subjects; PS, propensity score; Prog, prognostic score; LGM, logistic regression model; LRM, linear regression model; CART, classification and regression trees; RF, random forest; GBM, generalized boosted model; Average, model averaging prognostic score based on the prediction accuracy; ASMD, absolute standardized mean difference; KS, Kolmogorov-Smirnov

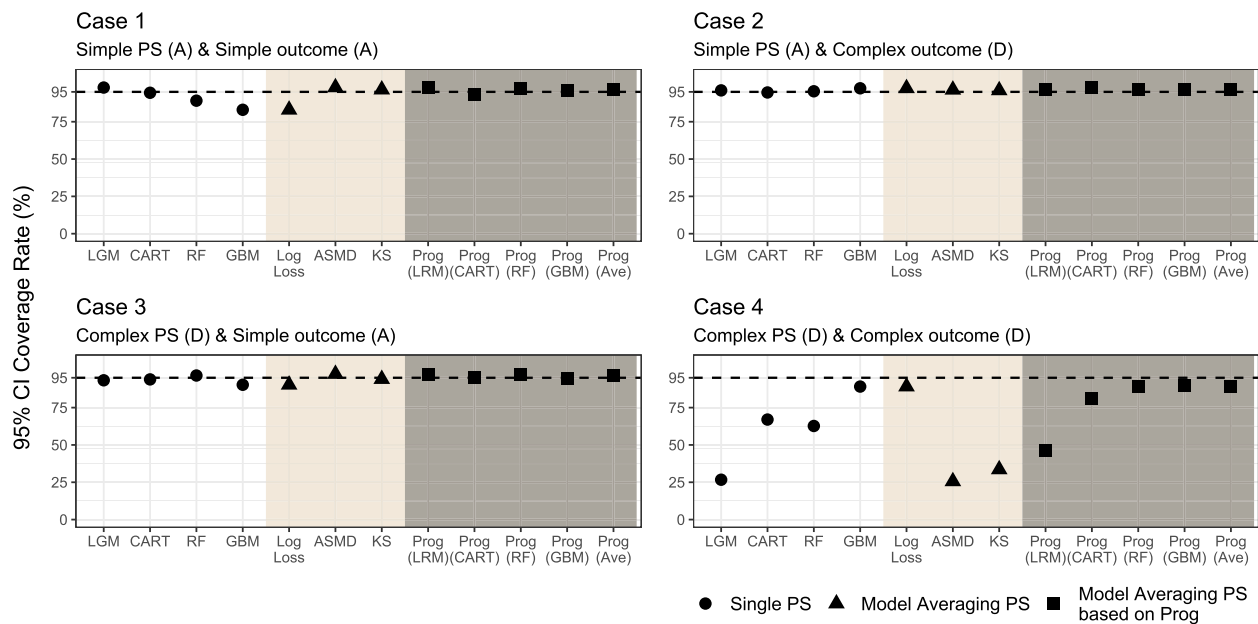


Fig. 3 95% CI Coverage Rate (%). Abbreviations: CI, confidence interval; PS, propensity score; Progn, prognostic score; LGM, logistic regression model; LRM, linear regression model; CART, classification and regression trees; RF, random forest; GBM, generalized boosted model; Average, model averaging prognostic score based on the prediction accuracy; ASMD, absolute standardized mean difference; KS, Kolmogorov-Smirnov

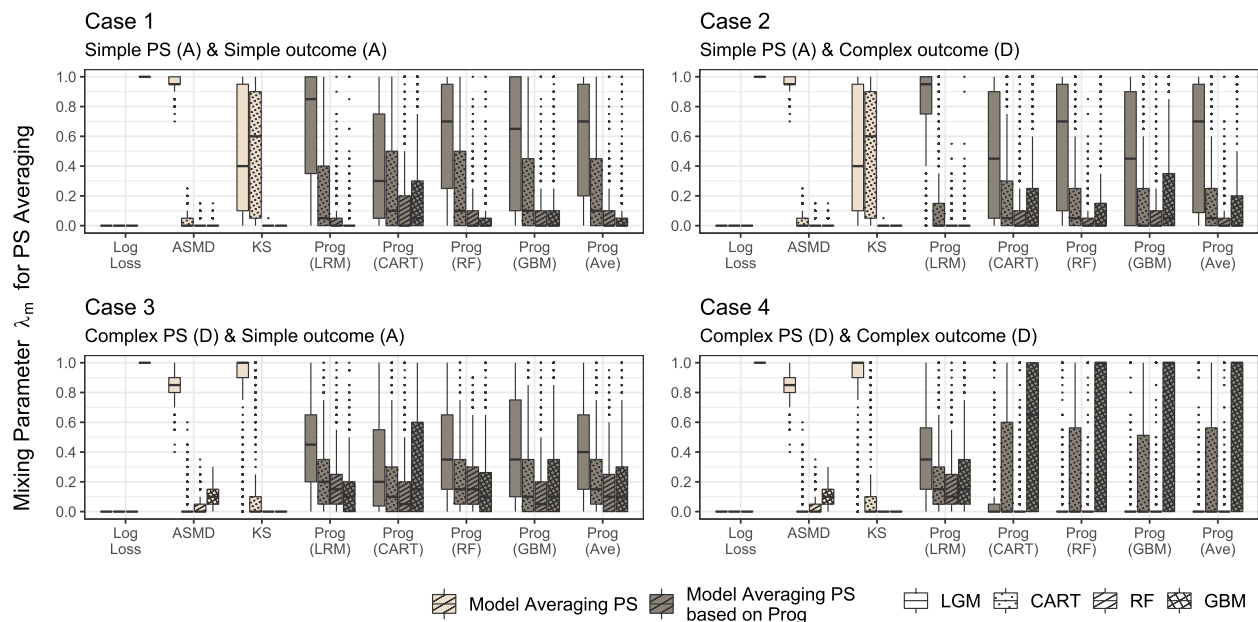


Fig. 4 Distributions of Mixing Parameter λ_m for PS Averaging. Abbreviations: PS, propensity score; Progn, prognostic score; LGM, logistic regression model; LRM, linear regression model; CART, classification and regression trees; RF, random forest; GBM, generalized boosted model; Average, model averaging prognostic score based on the prediction accuracy; ASMD, absolute standardized mean difference; KS, Kolmogorov-Smirnov

Results of single PS estimators

In scenarios where the true PS is simple (cases 1 and 2), the estimator with LGM-estimated PS exhibits the smallest absolute bias among single PS-based estimators. Conversely, in scenarios with a complex true PS (cases 3 and

4), the absolute bias of the LGM-estimated PS estimator is higher than the others. For case 3, the estimator using CART-estimated PS achieves the smallest bias. In case 4, the estimator employing GBM-estimated PS achieves the smallest bias.

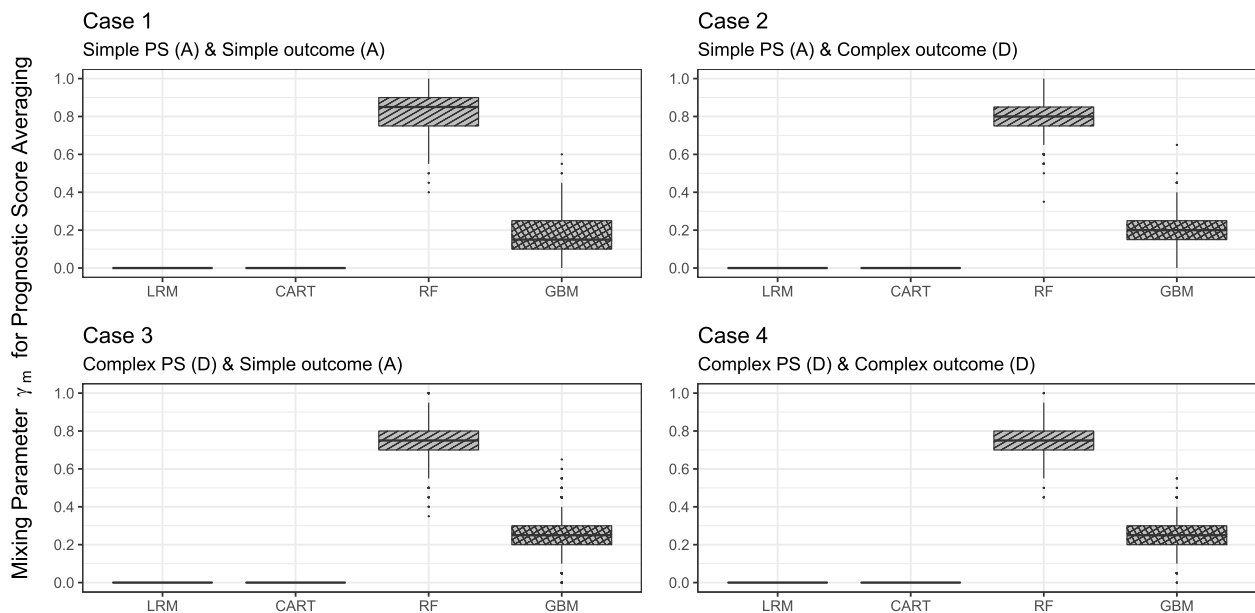


Fig. 5 Distributions of Mixing Parameter γ_m for Prognostic Score Averaging. Abbreviations: PS, propensity score; Prog, prognostic score; LRM, linear regression model; CART, classification and regression trees; RF, random forest; GBM, generalized boosted model

Across all cases, the estimator using RF-estimated PS achieves the smallest empirical standard error among the single estimators. Regarding RMSE, a composite performance measure of bias and variance, the lowest values occur in case 1 with LGM, in case 2 with RF and GBM, in case 3 with RF, and in case 4 with GBM.

The coverage rates of 95% CIs consistently exceed 80% in cases 1-3, regardless of PS estimating methods. However, in case 4, the coverage rates significantly drop, except for the estimator using GBM-estimated PS (26.7%, 67.0%, 62.7%, and 89.1% with LGM-, CART-, RF-, and GBM-estimated PS, respectively) (Fig. 3).

Results of existing model averaging approach

Across all iterations among all cases, PS estimation using GBM consistently produce the smallest Log Loss values (Supplemental Figure 1 and Fig. 4). Thus, the estimator employing model averaging PS, based on the predictive accuracy for treatment assignment, delivers results comparable to those of the single PS estimator from GBM-estimated PS (Fig. 4). Furthermore, the model averaging PS estimator using ASMD combines LGM and CART in cases 1 and 2, and LGM, RF, and GBM in cases 3 and 4. The estimator based on the KS statistic incorporates LGM and CART in all cases. In cases 1-3, the model averaging estimator based on Log Loss incurs a larger bias than those derived from model averaging estimators using covariate balance metrics (ASMD and KS statistic).

Conversely, in case 4, the estimator based on prediction accuracy results in the smallest bias.

Regarding RMSE, when the true outcome model is simple (cases 1 and 3), the model averaging approach using ASMD outperforms those based on other metrics. In contrast, for complex true outcome models (cases 2 and 4), the performance of the model averaging approach based on Log Loss surpasses those using covariate balance.

The coverage rates of 95% CIs for the model averaging estimator based on prediction performance remain above 80% across all cases. However, those for estimators using covariate balance significantly drop in case 4.

Results of proposed model averaging approach

The model averaging estimators using prognostic score balance exhibit similar performance across all methods for prognostic score estimation, except for case 4. In case 4, the estimator employing either RF- or GBM-estimated prognostic score as the balance metric achieves a smaller RMSE than the others. The prognostic score estimated with RF most accurately predicts the outcome values among untreated subjects, with GBM as the second most accurate across all cases (Fig. 5). Consequently, the model averaging prognostic score primarily combines RF and GBM estimates in approximately a 7:3 ratio (Fig. 5). When employing the model averaging prognostic score as the balance metric, the corresponding model averaging estimator demonstrates results closely aligned

with the best estimator in the prognostic score balanced approach.

The coverage rates stay near 95% across all cases, except in case 4, where using the LRM-estimated prognostic score led to a rate of 46%. This rate is still higher than the coverage rate achieved by the covariate balance-based model averaging approach (Fig. 3).

In all cases, the model averaging approach using the prognostic score as the balance metric yields lower or equivalent bias compared to both single PS-based approaches and existing model averaging approaches focused on prediction accuracy and covariate balances. Although the empirical standard error of the proposed model averaging approach is a little larger than the smallest standard error of the single PS approach, most of the RMSEs of the proposed estimators are smaller than those of the single PS approach. Moreover, compared to existing model averaging approaches, it is less affected by the function used as the balance metrics and exhibits less variability in results based on the context. Particularly, using the model averaging prognostic score can produce outcomes less reliant on the data-generating process.

Empirical example

We apply both the proposed and typical estimators to data from a retrospective observational study that evaluated the surgical modality of liver resection (laparoscopic or open) and its clinical outcomes in patients with hepatocellular carcinoma [18]. The ethics committee of Osaka City University approved this study (approval number: 2020-227). It encompasses 446 patients who underwent liver resections of 5 cm or less for hepatocellular carcinoma at Osaka City University Hospital from January

2018 to December 2019. Among this analytical population, 252 patients (56.5%) received laparoscopic liver resections.

We estimate the difference in the natural logarithmic volume of intraoperative blood loss (mL) between surgical modalities using similar estimators as described in the Simulation Experiment section. We consider laparoscopic liver resection as the treatment of interest and open liver resection as the reference treatment. Consequently, we estimate the prognostic score models using patients who underwent laparoscopic liver resection. Table 1 and Fig. 6A present the point estimates of the ATT and the 95% CI. In addition, Fig. 6B and C illustrate the mixing parameters λ_m and γ_m for PS model averaging and the prognostic score model averaging, respectively. The weighted balance metrics used for detecting the mixing parameters appear in Supplemental Figure 6. Across all cases, the point estimates and 95% CI results are less than 0, indicating that the intraoperative blood loss volume in the laparoscopic liver resection group is significantly lower than in the open liver resection group. However, each estimator yields slightly different results. Particularly, the estimate based on PS estimated by LGM deviates from those using PS estimated by other machine learning methods. Furthermore, the existing model averaging approach also produces variability in results, depending on the balance metrics used. These differences depend on which machine learning outcomes are more heavily reflected in each model averaging PS (Fig. 6B). In contrast, using the model averaging PS based on the prognostic score, results in similar outcomes. In this example, since RF is the most accurate in predicting the prognostic score, the model averaging prognostic

Table 1 Analysis results in example data

PS estimator	Balance metric	ATT estimate	Lower 95% CI	Upper 95% CI
LGM	-	-0.936	-1.464	-0.408
CART	-	-1.208	-1.546	-0.870
RF	-	-1.229	-1.518	-0.941
GBM	-	-1.146	-1.508	-0.784
Model Averaging PS	Log Loss	-1.229	-1.518	-0.941
	ASMD of Covariates	-1.116	-1.491	-0.741
	KS of Covariates	-0.980	-1.473	-0.486
Model Averaging PS based on Prog	ASMD of LRM-estimated Prog	-1.012	-1.484	-0.539
	ASMD of CART-estimated Prog	-1.012	-1.484	-0.539
	ASMD of RF-estimated Prog	-1.050	-1.486	-0.614
	ASMD of GBM-estimated Prog	-1.012	-1.484	-0.539
	ASMD of Model Averaging Prog	-1.050	-1.486	-0.614

Abbreviations: *ATT* average treatment effect on treated subjects, *CI* confidence interval, *PS* propensity score, *Prog* prognostic score, *LGM* logistic regression model, *LRM* linear regression model, *CART* classification and regression trees, *RF* random forest, *GBM* generalized boosted model, *ASMD* absolute standardized mean difference, *KS* Kolmogorov-Smirnov

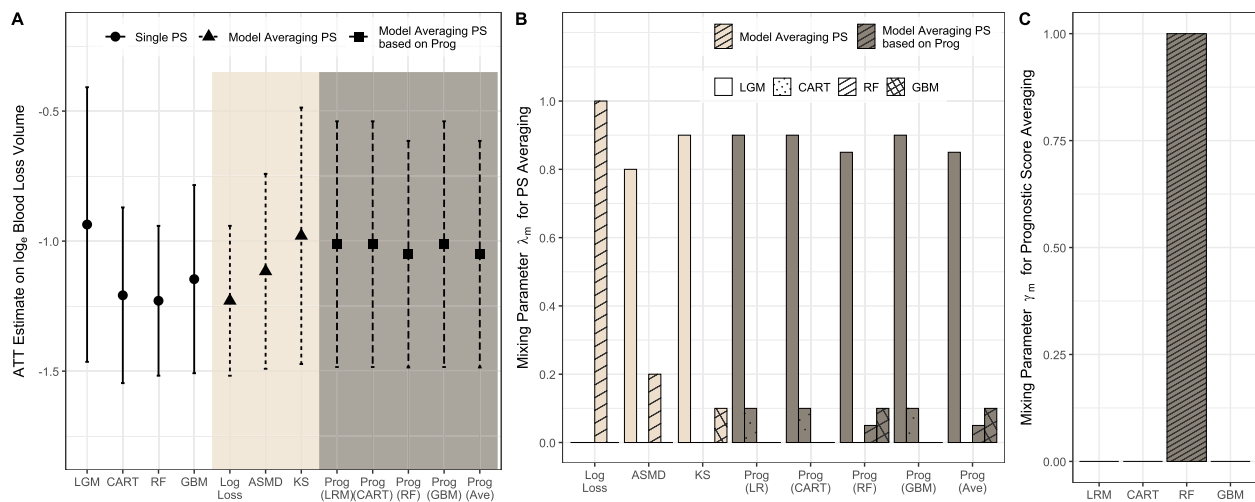


Fig. 6 Analysis Results in Example Data. Panel A shows the point estimates and 95% CIs of ATT on natural logarithmic blood loss volume. Panel B shows the mixing parameter λ_m for each PS estimating method. Panel C shows the mixing parameter γ_m for each prognostic score estimating method. Abbreviations: PS, propensity score; Prog, prognostic score; LGM, logistic regression model; LRM, linear regression model; CART, classification and regression trees; RF, random forest; GBM, generalized boosted model; Average, model averaging prognostic score based on the prediction accuracy; ASMD, absolute standardized mean difference; KS, Kolmogorov-Smirnov

score based on predictive accuracy coincides with results obtained using the RF-estimated prognostic score as the balance metric (Fig. 6C).

Discussion

We propose a model averaging approach for PS estimation based on prognostic score balance. To demonstrate and compare the proposed estimator's performance with typical estimators, we conduct simulation experiments under various data-generating processes. Additionally, we apply these estimators to an empirical example from a retrospective observational study.

In our simulation, among the estimators using a single PS model, the one employing LGM-estimated PS outperforms the others when the true PS model is simple. Conversely, when the true PS has a complex structure, the other estimators using machine learning-estimated PSs yield better performance than the one using LGM-estimated PS. This result aligns with the findings of other simulation studies [2, 3].

The estimators employing the model averaging approach, which minimizes the imbalance of the mean weighted ASMD and KS statistic, perform better than the estimator based on model averaging PS which minimizes predictive accuracy on treatment assignment in cases 1-3. These results are in line with assertions in several literatures suggesting a preference for selecting PS that optimizes covariate balance over predictive accuracy [10, 19]. On the other hand, when both the true PS and outcome models are complex, the model averaging approach based on prediction accuracy outperforms those focused

on covariate balance. This result suggests that the optimal PS model varies depending on the complexity of PS model and the relationship between covariates and the outcome.

Our proposed model averaging approaches based on prognostic score balance result in biases that are similar to or fewer than those in the existing model averaging models proposed by Xie et al. in most cases. Furthermore, the performances are not sensitive to the methods used for prognostic score estimation in cases 1-3, aligning with the results shown in the experiment by Stuart et al. [7]. On the other hand, when both PS and outcome models have complex structures, the model averaging approach based on RF and GBM-estimated prognostic scores, known for their high predictive accuracy in outcome values, achieves better performance. Since prognostic scores are derived from variables directly related to the outcome, using prognostic scores with high predictive accuracy as a balancing metric between treatment groups instead of covariates supports the idea that this method can reduce bias and variance in estimated treatment effects. Furthermore, when using the model averaging prognostic score based on their predictive accuracy as the balance metrics, the estimator leads to better performance compared to the other estimators across all cases.

The trends observed in the simulation also appear in the analysis of empirical example data. The estimator using model averaging PS based on the balance of prognostic scores yields more stable results compared to the single PS estimators and the existing model averaging

PS estimators. Similarly, the model averaging PS estimator using the model averaging prognostic score as a balance metric produces comparable results. These results further indicate that our proposed estimator reduces the risk of generating estimates with extreme errors dependent on specific situations.

This study has several limitations. One important consideration in our study is the adherence to the principle of separation between the design and analysis phases [19, 20]. Existing approaches consider covariate balance without accessing outcome information, similar to randomized experiments, which ensures the separation principle between design and analysis. In contrast, our proposed approach using prognostic scores incorporates outcome information for model selection, limiting its use to study designs where the outcomes are predetermined. This compromises the independence of study design and analysis method selection, raising concerns that differences in chosen designs may affect the analysis results. However, the prognostic score model estimation is based on data from untreated subjects, and the prognostic score is used only for PS selection and not directly for treatment effect estimation, so the principle of separation may not be entirely violated [7]. This concern can be mitigated by separating the data used for the model estimation from the data used to fit the estimated model and obtain treatment effect estimates, namely, the data-splitting approach. Moreover, covariate information homogenized by random treatment assignment can be used to analyze multiple outcomes. This is one of the significances of separating design and analysis. Existing approaches ensure covariate balance independent of outcomes, but our proposed approach requires estimating a prognostic score for each outcome of interest. Therefore, in our experiments, we focused only on one continuous primary outcome. Future research needs to extend our method to be applicable when analyzing multiple outcomes. Furthermore, our simulation setting, basically inspired by Setoguchi et al., is limited in its simplicity, which reduces its applicability to real-world data analysis contexts. Naimi, Mishler, and Kennedy (2023), along with Balzer and Westling (2023), have suggested that using the IPW estimator with machine learning for PS estimation can produce invalid variances and biased estimates when the true nuisance functions are complex [21, 22]. The unreliability increases as the complexity and dimensionality of the model estimated by machine learning rise, due to a slow convergence rate and regularization bias which can be referred to as an omitted-variable bias [23]. To mitigate these effects, they recommend using doubly robust estimators and estimation approaches employing data-splitting. The

prognostic score-based averaging PS model we propose can also be utilized as part of the nuisance function estimates used in doubly robust estimation. Additionally, data-splitting can be easily applied to our proposed method, which not only reduces regularization bias but also supports the principle of separating study design and analysis mentioned earlier. Moreover, they emphasize the importance of considering a wider range of machine learning candidates for an ensemble estimation whereas our study employs limited machine learning due to computing resource constraints. When applying our proposed estimator to actual data analysis, it is recommended to consider the limitations of our study and to combine the proposed estimates with the approaches mentioned above, depending on the complexity and dimensionality of the data being analyzed.

In conclusion, the prognostic score-based model averaging approach for estimating PS can outperform existing model averaging methods. In particular, the estimator using the model averaging prognostic score as a balance metric can produce more robust estimates, making it recommendable for a wide variety of situations. When applying the proposed method to real-world data, it is important to use it in conjunction with techniques that mitigate issues arising from the complexity and high dimensionality of the data.

Abbreviations

PS	Propensity score
ATE	Average treatment effect
ATT	Average treatment effect on the treated subjects
IPW	Inverse probability weighted estimator
LGM	Logistic regression model
LRM	Linear regression model
CART	Classification and regression trees
RF	Random forest
GBM	Generalized boosted model
ASMD	Absolute standardized mean difference
KS	Kolmogorov-Smirnov
MSE	Mean squared error

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12874-024-02350-y>.

Supplementary Material 1.

Acknowledgements

The authors express their gratitude to Daisuke Shirai and Hiroji Shinkawa for collecting the clinical data used in the empirical example.

Authors' contributions

DK contributed to the conceptualization, conducted the simulations, analyzed the data, and drafted the manuscript for this study. ES and AS provided advice on the simulation and data analysis. All authors read and approved the final manuscript.

Funding

This work is supported by JSPS KAKENHI (JP23K17245).

Availability of data and materials

The R code for generating the simulation data and obtaining the treatment effect estimates is provided in Supplemental Text 1. The R code used for the simulations is heavily based on the publicly available code by Lee et al. and Chernozhukov et al. [24, 25]. The dataset used in the EMPIRICAL EXAMPLE section is not publicly available because it is currently being used in another ongoing project. However, we invite inquiries about the data's details or future use to be directed to the corresponding author.

Declarations

Ethics approval and consent to participate

As documented in the EMPIRICAL EXAMPLE section, the example data to which we applied our proposed analytical method received approval from the Ethics Committee of Osaka City University (approval number: 2020-227).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Center for Mathematical and Data Science, Kobe University, 1-1 Rokko-dai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan. ²Department of Medical Statistics, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan. ³Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA. ⁴Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

Received: 25 March 2024 Accepted: 23 September 2024

Published online: 03 October 2024

References

- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *1983*;70(1):41–55.
- Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiol Drug Saf.* 2008;17(6):546–55.
- Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med.* 2010;29(3):337–46.
- Pirracchio R, Petersen ML, Van Der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol.* 2015;181(2):108–19.
- Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf.* 2011;20(3):317–20.
- Xie Y, Zhu Y, Cotton CA, Wu P. A model averaging approach for estimating propensity scores by optimizing balance. *Stat Methods Med Res.* 2019;28(1):84–101.
- Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures for propensity score methods in comparative effectiveness research. *J Clin Epidemiol.* 2013;66(8):S84–S90.e1.
- Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res.* 2009;18(1):67–80.
- Hansen BB, Hansen BB. The prognostic analogue of the propensity score. *Biometrika.* 2008;95(2):481–8.
- Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol.* 2006;163(12):1149–56.
- Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology - an empirical illustration. *Pharmacoepidemiol Drug Saf.* 2011;20(6):551–9.
- Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Practice of epidemiology effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol.* 2011;174(11):1213–22.
- Shortreed SM, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics.* 2017;73(4):1111–22.
- Kabata D, Shintani M. Variable selection in double/debiased machine learning for causal inference: an outcome-adaptive approach. *Commun Stat Simul Comput.* 2021;52(12):5880–93.
- Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med.* 2004;23(19):2937–60.
- Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med.* 2015;34(28):3661–79.
- Van Der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol.* 2007;6(1). <https://doi.org/10.2202/1544-6115.1309>.
- Shirai D, Shinkawa H, Kabata D, Takemura S, Tanaka S, Amano R, et al. Laparoscopic liver resection reduces postoperative infection in patients with hepatocellular carcinoma: a propensity score-based analysis. *Surg Endosc.* 2022;36(12):9194–203.
- Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf.* 2004;13(12):855–7.
- Rosenbaum PR. Various Practical Issues in Matching. In: Rosenbaum PR, editor. *Design of Observational Studies*. New York: Springer New York; 2010. p. 187–95.
- Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *Am J Epidemiol.* 2023;192(9):1536–44.
- Balzer LB, Westling T. Invited commentary: demystifying statistical inference when using machine learning in causal research. *Am J Epidemiol.* 2023;192(9):1545–9.
- Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J.* 2018;21(1):C1–68.
- Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS ONE.* 2011;6(3):e18174.
- Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. Double/debiased/neyman machine learning of treatment effects. *Am Econ Rev.* 2017;107(5):261–5.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.