Received: 3 May 2021 | Revised: 26 November 2021 | Accepted: 1 December 2021

DOI: 10.1002/sim.9289

RESEARCH ARTICLE

Statistics in Medicine WILEY

Practical recommendations on double score matching for estimating causal effects

Yunshu Zhang[1] | **Shu Yang**[1] | **Wenyu Ye**[2] | **Douglas E. Faries**[2] | **Ilya Lipkovich**[2] | **Zbigniew Kadziola**[2]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

[2]Eli Lilly and Company, Indianapolis, Indiana, USA

Correspondence
Yunshu Zhang, Department of Statistics, North Carolina State University, SAS Hall, 2311 Stinson Dr, Raleigh, NC 27607, USA.
Email: yzhan234@ncsu.edu

Funding information
National Institute of Aging, Grant/Award Number: 1R01AG066883; National Institute of Environmental Health Sciences, Grant/Award Number: 1R01ES031651; National Science Foundation, Grant/Award Number: DMS 1811245

Unlike in randomized clinical trials (RCTs), confounding control is critical for estimating the causal effects from observational studies due to the lack of treatment randomization. Under the unconfoundedness assumption, matching methods are popular because they can be used to emulate an RCT that is hidden in the observational study. To ensure the key assumption hold, the effort is often made to collect a large number of possible confounders, rendering dimension reduction imperative in matching. Three matching schemes based on the propensity score (PSM), prognostic score (PGM), and double score (DSM, ie, the collection of the first two scores) have been proposed in the literature. However, a comprehensive comparison is lacking among the three matching schemes and has not made inroads into the best practices including variable selection, choice of caliper, and replacement. In this article, we explore the statistical and numerical properties of PSM, PGM, and DSM via extensive simulations. Our study supports that DSM performs favorably with, if not better than, the two single score matching in terms of bias and variance. In particular, DSM is doubly robust in the sense that the matching estimator is consistent requiring either the propensity score model or the prognostic score model is correctly specified. Variable selection on the propensity score model and matching with replacement is suggested for DSM, and we illustrate the recommendations with comprehensive simulation studies. An R package is available at https://github.com/Yunshu7/dsmatch.

KEYWORDS
average treatment effect on the treated, causal inference, double robustness, prognostic score, propensity score

1 | INTRODUCTION

Randomized clinical trials (RCTs) are the touchstone for treatment effect evaluation. By trial design, treatment randomization guarantees that treatment groups are comparable and thus bias can be minimized to the extent possible. However, in practice, it may be infeasible to conduct an RCT due to financial, logistic, or ethical reasons. In these settings, comparative analyses using observational data may be of particular value. Unlike RCTs, confounding control is critical for estimating the causal effects from observational studies due to the lack of treatment randomization. Under

the unconfoundedness assumption, that is, all pre-treatment variables that are predictors of treatment and outcome are observed, matching methods[1,2] can be used to emulate an RCT that is hidden in the observational study. Researchers from various disciplines have expanded this field in both theory and practice for decades; see Reference 2 for a comprehensive review. To ensure the key assumptions hold, the effort is often made to collect a large number of possible confounders, rendering dimension reduction imperative in matching.

In their seminal paper, Rosenbaum and Rubin[3] demonstrated the vital role of the propensity score, which is defined as the conditional probability of receiving treatment given the confounders, as a balancing score. The key implication is that matching on the scalar propensity score can dramatically reduce the confounding bias. Since then, propensity score matching (PSM) has been the most common method in the industry. However, recently, PSM has been criticized to be ineffective in that it attempts to emulate a completely randomized trial,[4] which is rarely if at all implemented in practice.

The prognostic score is an important alternative score summarizing covariate correlations with the outcomes. It is also known as the disease risk score in the epidemiology literature with a rich history.[5] Its fundamental theory as a balancing score was established by Hansen[6] in 2008, and he suggested various ways to utilize prognostic score in matching, called prognostic score matching (PGM). PGM balances the disease risk between the treatment groups and thus attempts to emulate a special blocked randomized trial, where the blocks are formed by the risk levels. The reason we call it a special trial is that not all the covariates are balanced within blocks as in standard blocked randomized design. Only prognostic factors that are highly related to the outcomes are adjusted for. More importantly, the potential outcomes are balanced within blocks, which is our ultimate goal in a trial. By not adjusting for unimportant variables that are not related to the outcomes, the special design may obtain higher efficiency. Wyss et al[7] used simulations and an empirical example to illustrate the superior efficiency of PGM compared to PSM when the propensity score distributions are separated. Another notable advantage of PGM is that it is less sensitive than PSM to the practical violation of the overlap assumption. This is because treatment selection often shifts the propensity score distribution more dramatically than the prognostic score distribution between the treatment groups (see, eg, Figure 3). As a result, the matching rate in PGM can be larger than that in PSM. However, this may not be true when the prognostic score is highly correlated with the propensity score. In this case, the prognostic score cannot provide any additional information compared to the propensity score. Thus, it is often reasonable to check the correlation between the two scores.[8] Moreover, Stuart et al[9] showed empirically the positive correlation between the prognostic score and the bias of the treatment effect estimator and thus the prognostic score is useful for balance check after matching. Nguyen and Debray extended the use of the prognostic score[10] to the case of general treatment regimes.

Combining the propensity score and prognostic score is indeed a sensible alternative to form a balancing score. Leacy and Stuart[11] were the first in the statistical and epidemiological literature to assess the performance of jointly matching or stratifying on both the propensity score and prognostic score via simulation, although this idea was first raised by Hansen[12] in a technical report. In addition to the full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores and the full matching on the estimated prognostic propensity score (ie, the propensity score predicted by the estimated prognostic score) within propensity score calipers, their paper also included subclassification on the two scores. As a result, although PGM achieves the best performance when the prognostic score model is correctly specified, full matching using the Mahalanobis distance combining the estimated propensity and prognostic scores is more robust to model misspecifications, while its performance is similar to PGM. Antonelli et al[13] later established the double robustness and convergence rate of the double score estimator of the average treatment effect (ATE). Yang and Zhang[14] coined the term "double score matching" (DSM) and derived the asymptotic distribution of the double score estimator for the average and quantile treatment effects. Both studies underscore the advantage of DSM being doubly robust against model misspecification of either the propensity score model or the prognostic score model. Besides matching, the propensity score and the outcome information are used in the augmented inverse probability weighting estimator that achieves semiparametric efficiency bound and double robustness. Hu et al[15] utilized the double score in the context of nonparametric estimation.

In short, propensity score, prognostic score, and double score are useful for dimension reduction and reducing confounding bias, resulting in three matching schemes. However, a comprehensive comparison among the three matching schemes is lacking and research towards the best practices including variable selection and choices of whether using caliper and replacement is still needed. First, an important topic is variable selection in the propensity score and prognostic score. In fact, model fitting has been an important and challenging issue in PSM. Brookhart et al[16] used simulations to suggest using covariates related to outcome and not using instrumental variables (IVs) to construct the propensity score model for propensity score weighting. Yang et al[17] recommended selecting all variables that are predictive of either treatment or outcome for robustness consideration. De Luna et al[18] proposed an algorithm to select the minimal sets of

TABLE 1 Comparison among the three matching estimators

Matching scheme	The emulating trial design	Model specification requirement	Overlap requirement	Optimal configuration*			Performance under optimal configuration*
				Variable selection	Replacement	Caliper	
PSM	Completely randomized experiment	The propensity score model	Sufficient overlap of the propensity score distribution	Yes	Yes	Yes	Lower matching rate and large variance
PGM	Special blocked randomized experiment	The prognostic score model	Sufficient overlap of the prognostic score distribution	No	Yes	Insensitive	High matching rate and small variance
DSM	Hybrid randomized experiment	Either the propensity score or prognostic score model (double robustness)	Sufficient overlap of either the propensity score or prognostic score distribution	Yes for the propensity score model	Yes	Insensitive	High matching rate and small variance

Note: Results with * are based on the correct model specifications.

covariates. Myers et al[19] reported similar findings in PSM but prioritized minimizing unmeasured confounding when selecting variables, even at the risk of conditioning on IVs. At the same time, Pearl[20] showed that the rate of bias amplification from IVs may be faster than the rate of bias reduction. Some researchers suggested that IVs may be less detrimental if the prognostic score is included in matching.[5] Most existing recommendations are confined to PSM, and very few studies have investigated these issues for PGM and DSM.

Second, matching constraints are also important issues in matching methods, including whether to include caliper and whether to match with or without replacement. In PSM, a caliper of 0.25 standard deviation of linear propensity score was generally suggested by Rosenbaum and Rubin.[21] Other researchers used Monte Carlo methods to find the optimal caliper width equal to 0.2 of the standard deviation in some special circumstances.[22,23] Matching without replacement is claimed to reduce variance since each control sample is used only once.[24] But when the control group is not large enough, bias can often be decreased by matching with replacement.[25] However, these recommendations are restricted to PSM and may not extend to PGM and DSM given that the prognostic score distribution is less affected by treatment assignment than the propensity score distribution.

In this article, we explore the statistical and numerical properties of the three score-based matching methods via extensive simulations. Our study supports the conclusion that DSM performs comparably with, if not better than, the other two single score matching. Table 1 summarizes the key features for comparison. Importantly, by linking the matching scheme and trial design, we show that DSM emulates a hybrid design of complete randomization and blocked randomization that incorporates the blocking benefit of PGM while retaining the balancing guarantee of PSM. We also provide bolts and nuts for DSM and illustrate the recommendations with comprehensive simulation studies. In essence, we propose the following steps for DSM to achieve its best performance in terms of bias and variance.

Step 1. At the variable selection stage, fit a prognostic score model with penalization (eg, LASSO) to select all prognostic variables.

Step 2. Fit the propensity score model restricted to the selected prognostic variables and fit the prognostic score model with all the covariates.

Step 3. Calculate the double score and assess the overlap of the propensity score and the prognostic score, or alternatively, assess the matching rate after implementing matching based on the double score in the following step. We require sufficient overlap for at least one score or a large matching rate to apply DSM.

Step 4. Carry out matching with replacement based on the estimated double score and calculate the ATE estimator.

Thanks to the variable selection strategy, DSM can achieve the best efficiency across different settings in the simulation studies, in contrast to Leacy and Stuart's result that PGM performed best when models were correctly specified. Also, because all the variables are kept in the prognostic score model, the estimator will not suffer from confounding bias. This is a special advantage of DSM over PSM and PGM in consideration of variable selection. The double robustness property of DSM is also verified by extensive simulations across various scenarios. However, it is worth noting that all the above conclusions are based on limited simulation studies where covariates and outcomes are normally distributed continuous variables. Readers should be cautious not to apply these results to other scenarios before a general theoretical proof is derived.

The remaining sections of this article proceed as follows. Section 2 introduces the three matching methods: PSM, PGM, and DSM. Section 3 focuses on model selection and can be divided into two parts: the first part uses a hypothetical simulating example to illustrate the performance of different variable selection strategies; the second part turns to a more realistic simulation setting using the REFLECTIONS dataset.[26] Because of the complex correlation in the covariate set, LASSO is applied to select variables into the outcome models. Simulation setup and detailed results are both presented in this part. Section 4 presents the simulation results regarding the choice of caliper and replacement. Section 5 compares the three matching estimators based on their best configurations of variable selection strategy and choice of caliper and replacement. Section 6 concludes the article and discusses possible directions of future work. Derivations of the asymptotic results for the three matching estimators and the comparison of their theoretical efficiency are included in the Appendix.

2 | METHODOLOGY: MATCHING ESTIMATORS

We follow the potential outcome framework to formulate causal effects of treatment.[24] Let X_i be the set of covariates, A_i be the treatment indicator, and Y_i be the observed outcome for unit $i = 1, \dots, n$. Let $Y_i(a)$ be the potential outcome had unit i been given treatment a, where $a = 1$ is the treatment of interest and $a = 0$ is the control group. It is assumed that $\{X_i, A_i, Y_i(0), Y_i(1)\}, i = 1, \dots, n$ are independent and identically distributed. We intend to estimate the ATE on the treated $\tau_{\text{ATT}} = E\{Y(1) - Y(0) | A = 1\}$.

We use matching to impute missing potential outcomes in causal inference. The main intuition is to find the closest subject in the control group for each individual in the treatment group. Note that here we restrict our analysis within the estimation of the ATE on the treated (ATT) and one-to-one matching. Thus, the key step in matching is to define a proper distance measure to determine "closeness." PSM, PGM, and DSM are three variants with different measures of distance, and we will introduce them in detail in this section. Besides, calipers can be incorporated into the definition of distance so that subjects who cannot find a good match will be excluded from the analysis. Moreover, matching with or without replacement determines whether a control individual can be matched multiple times or not.

2.1 | Propensity score matching

The propensity score is defined as the probability of receiving treatment given certain values of covariates:[3]

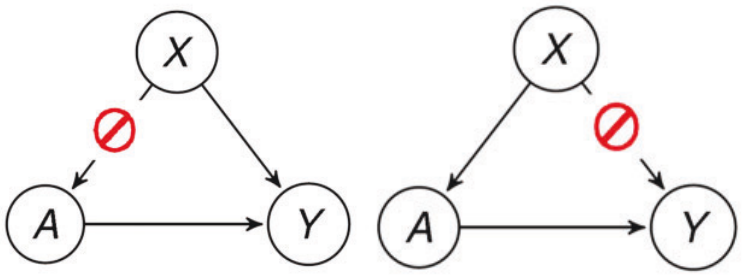
$$e(X) = E(A | X) = P(A = 1 | X).$$

Here we restrict our analysis to binary treatment scenarios. As illustrated in Figure 1A, the path between covariates and treatment is blocked by the propensity score. Under positivity and no unmeasured confounder assumptions (stated formally as Assumptions 1 and 2 later in the Appendix), the propensity score is a balancing score conditional on which the potential outcomes and treatment assignment are independent:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid e(X).$$

This implies that the estimand ATT can be written as:

$$\begin{aligned}\tau_{\text{ATT}} &= E\{Y(1) - Y(0) | A = 1\} \\ &= E[E(Y | A = 1) - E\{Y | A = 0, e(X)\} | A = 1].\end{aligned}$$



(A) DAG of PSM **(B)** DAG of PGM

FIGURE 1 DAGs to illustrate the different intuitions behind PSM and PGM. PSM blocks the path from covariates to treatment, while PGM blocks the path from covariates to outcome

This ensures that we can use PSM to estimate the ATT. We define the distance D_{ij} between individuals i and j for matching based on the linear propensity score:

$$D_{ij} = \left|\text{logit}\left\{e\left(X_i\right)\right\} - \text{logit}\left\{e\left(X_j\right)\right\}\right|. \tag{1}$$

The linear version is used because the bias may be effectively reduced.[21,27,28] Later on, we will use the "propensity score" to refer to either the probability version or the linear version. Suppose there are n units and individuals indexed by $1, \ldots, n_1$ are from the treatment group. For each treated unit i, we denote J_i as the index of the closest control unit in the sense of distance defined by the propensity score:

$$J_i = \underset{j=n_1+1,\ldots,n}{\arg\min}\, D_{ij}. \tag{2}$$

As a result, the ATT can be estimated by

$$\hat{\tau}_{\text{ATT}} = n_1^{-1}\sum_{i=1}^{n_1}\left(Y_i - Y_{J_i}\right). \tag{3}$$

To improve matching quality, we can incorporate a caliper into the definition of matching index J_i:

$$J_i = \begin{cases} \underset{j=n_1+1,\ldots,n}{\arg\min}\, D_{ij}, & \text{if } \underset{j=n_1+1,\ldots,n}{\min} D_{ij} \le c, \\ 0, & \text{if } \underset{j=n_1+1,\ldots,n}{\min} D_{ij} > c, \end{cases} \tag{4}$$

where c is the preset caliper. This excludes treated units that are dissimilar from anyone in the control group. Let $\mathcal{J}$ be the set of treated indices that will be included in matching:

$$\mathcal{J} = \{i : J_i \neq 0, i = 1, \ldots, n_1\}.$$

Then the ATT can be estimated by

$$\hat{\tau}_{\text{ATT}} = |\mathcal{J}|^{-1}\sum_{i\in\mathcal{J}}\left(Y_i - Y_{J_i}\right). \tag{5}$$

Note that the above estimators are based on matching with replacement. When matching without a caliper and without replacement, J_i is chosen from the remaining control indices:

$$J_i = \underset{j\in\left\{n_1+1,\ldots,n\right\}\setminus\left(\bigcup_{i'=1}^{i-1}\left\{J_{i'}\right\}\right)}{\arg\min} D_{ij}. \tag{6}$$

When matching with a caliper and without replacement, the definition of J_i can be adapted in a similar way:

$$J_i = \begin{cases} \underset{j \in \{n_1+1,\ldots,n\} \setminus \left(\bigcup_{i'=1}^{i-1} \{J_{i'}\} \right)}{\arg\min} D_{ij}, & \text{if } \min_{j \in \{n_1+1,\ldots,n\} \setminus \left(\bigcup_{i'=1}^{i-1} \{J_{i'}\} \right)} D_{ij} \leq c, \\ 0, & \text{if } \min_{j \in \{n_1+1,\ldots,n\} \setminus \left(\bigcup_{i'=1}^{i-1} \{J_{i'}\} \right)} D_{ij} > c. \end{cases} \tag{7}$$

Note that in practice we do not have access to the true propensity score. Thus, the estimated propensity score $\hat{e}(X)$ will replace $e(X)$ in (1). In the simulation part of this article, a logistic model is always used to fit the propensity score, but the set of variables may vary depending on our model selection strategy.

2.2 | Prognostic score matching

The prognostic score $\Psi(X)$ is formally defined by Hansen[6] as a balancing score in the sense that $Y(0) \perp\!\!\!\perp X | \Psi(X)$. We illustrate the prognostic score using the following examples.

Example 1. If $Y(0)$ follows a generalized linear model with mean $\mu_0(X) = X^T\beta_0$ and constant variance, then $\Psi(X) = E(Y(0)|X) = X^T\beta_0$.

Example 2. If $Y(0)$ follows a location-shift family $f_0\{y - \mu_0(X)\}$, then $\Psi(X) = \mu_0(X)$.

As illustrated in Figure 1B, the path between covariates and outcome is blocked by the prognostic score. Hansen showed that if there is no hidden bias, treatment ignorability holds by conditioning on the prognostic score:

$$Y(0) \perp\!\!\!\perp A \mid \Psi(X).$$

This implies that the ATT can be estimated by matching via the prognostic score:

$$\begin{aligned} \tau_{\text{ATT}} &= E\{Y(1) - Y(0) \mid A = 1\} \\ &= E[E(Y \mid A = 1) - E\{Y \mid A = 0, \Psi(X)\} \mid A = 1]. \end{aligned}$$

Similarly, we define the distance D_{ij} between individuals i and j for matching based on the prognostic score:

$$D_{ij} = \left\| \Psi(X_i) - \Psi\left(X_j\right) \right\|. \tag{8}$$

Then we can estimate the ATT via formulas (2) to (7) depending on our choice of caliper and replacement. Similarly, an estimated prognostic score $\hat{\Psi}(X)$ will replace $\Psi(X)$ in (8), and a generalized linear model is used throughout this article as in Example 1. Note that only the control group is used in the model fitting step but prognostic scores of units from both groups are required to be estimated.

2.3 | Double score matching

Antonelli et al[13] showed that the double score as the combination of the propensity score and prognostic score is also a balancing score, and that conditioning on the double score deconfounds the potential outcomes from the treatment assignment:

$$Y(0) \perp\!\!\!\perp A \mid e(X), \Psi(X).$$

Note that this result holds even if only one score is correctly specified. This is the basis of the double robustness property of the DSM estimator. As a result, we can estimate the ATT by matching on the double score:

$$\begin{aligned}\tau_{\text{ATT}} &= E\{Y(1) - Y(0) \mid A = 1\} \\ &= E[E(Y \mid A = 1) - E\{Y \mid A = 0, e(X), \Psi(X)\} \mid A = 1].\end{aligned}$$

To be specific, the distance metric is defined as the Mahalanobis distance combining propensity score and prognostic score:

$$D_{ij} = \begin{pmatrix} \text{logit}\{e(X_i)\} - \text{logit}\{e(X_j)\} \\ \Psi(X_i) - \Psi(X_j) \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \text{logit}\{e(X_i)\} - \text{logit}\{e(X_j)\} \\ \Psi(X_i) - \Psi(X_j) \end{pmatrix}, \tag{9}$$

where Σ is the covariance matrix of $\left(\text{logit}\{e(X)\}, \Psi(X)\right)^T$. Mahalanobis distance removes the imbalance between the scales of two scores.

Formulas (2) to (7) can still be used to estimate the ATT, but Yang and Zhang[14] recommended to include a bias correction term into the estimator:

$$\hat{\tau}_{\text{ATT}} = n_1^{-1} \sum_{i=1}^{n_1} \left(\left(Y_i - Y_{J_i}\right) - \left[\hat{\mu}_0\{e(X_i), \Psi(X_i)\} - \hat{\mu}_0\left\{e\left(X_{J_i}\right), \Psi\left(X_{J_i}\right)\right\}\right] \right), \tag{10}$$

where $\mu_0\{e(X_i), \Psi(X_i)\} = E\{Y(0) \mid e(X_i), \Psi(X_i)\}$ and we can obtain its rough estimate by the method of sieves. Note that (10) only works when caliper is not included. When matching with a caliper, we can simply calculate the bias correction term using individuals who remained in the matching set as (5). Although Yang and Zhang showed that this bias is asymptotically negligible, correcting for bias may increase finite sample performance in practice.

There is another special issue about DSM when a caliper is included. Hansen[6] suggested using an ordinary propensity score caliper and matching on the prognostic propensity score. Leacy and Stuart[11] found that this is not better than matching based on the Mahalanobis distance of the two scores. Alternatively, we would like to incorporate a caliper in both matching by the propensity score and the prognostic score, in a procedure similar to PSM and PGM. However, we do not need to require both scores to be close enough. The intuition is that individuals within a matching pair are comparable if at least one of the scores satisfy the caliper constraint. To be specific, the matching index J_i in (4) changes its definition to:

$$J_i = \begin{cases} \underset{j=n_1+1,\ldots,n}{\arg\min}\, D_{ij}, & \text{if } \underset{j=n_1+1,\ldots,n}{\min} \min\left\{ \left|\text{logit}\{e(X_i)\} - \text{logit}\{e(X_i)\}\right|, \left|\Psi(X_i) - \Psi\left(X_j\right)\right| \right\} \le c, \\ 0, & \text{if } \underset{j=n_1+1,\ldots,n}{\min} \min\left\{ \left|\text{logit}\{e(X_i)\} - \text{logit}\{e(X_i)\}\right|, \left|\Psi(X_i) - \Psi\left(X_j\right)\right| \right\} > c. \end{cases}$$

When matching with replacement, the matching index J_i in (7) can be adapted similarly. As a result, the matching rate should be higher than both PSM and PGM. More importantly, overlapping assumptions are also relaxed. For units outside the overlapping region of the propensity score, they can still find matching pairs if they are in the overlapping region of the prognostic score. This is illustrated by the increased matching rate in the simulation results, and we hope to complete the theoretical proof in the future.

2.4 | Summary of the three matching methods

We provide some insights by connecting the three matching methods with randomized trials; see also Table 1. King and Nielsen[4] linked PSM with a completely randomized experiment. On the other hand, PGM mimics a more efficient fully blocked randomized experiment with a special design where only potential outcomes and prognostic factors are adjusted for. This explains why PGM may have a smaller variance than PSM when models are correctly specified. Meanwhile, DSM approximates a hybrid randomized experiment, which can be more efficient with the correctly specified prognostic score. When the outcome model is misspecified but the propensity score model is correct, DSM is still valid because it inherits the unbiasedness property from the complete randomized experiment.

So far, we have introduced the three types of estimators from PSM, PGM, and DSM, as well as their variants depending on different choices of caliper and replacement. It is natural to derive the theoretical results for these estimators. In the Appendix, we provide some asymptotic results for estimating ATT. These are based on some simplifications of the problem, following Abadie and Imbens.[29] First, we only consider matching with replacement and without a caliper. Second, the asymptotic results are established when the coefficients in the propensity score and the prognostic score models are known. Under these assumptions, the theory justifies the high efficiency of PGM compared to DSM, which invokes the importance of involving variable selection algorithms into DSM. However, these two assumptions we made may oversimplify the problem as the errors from the estimation of coefficients are not addressed, which may have a significant impact especially in high dimensional settings. Thus, the theory can only offer suggestions instead of guarantees. Meanwhile, when considering additional factors such as variable selection, it is difficult to make comparisons theoretically. In the following three sections, we will use simulations to study the effect of variable selection strategy, matching constraints, and distance metric on the matching estimator and make some recommendations based on the results.

3 | VARIABLE SELECTION IN MATCHING

In this section, we would like to address the questions: how variable selection will change the performance of each matching estimator and what kind of variable selection strategy we should use. To better answer these questions, we will first introduce an illustrating example that helps us understand the categories of variables and gives us guidance about the appropriate variable selection strategies. Then we will turn to a more comprehensive simulation study based on a more realistic dataset.

3.1 | Importance and strategy of variable selection: An illustrating example

3.1.1 | Simulation setup

We generated covariates for 3000 subjects from a 16-dimensional multivariate normal distribution, where the mean vector was 0 and the covariance matrix was a 16×16 identity matrix. The covariate distribution was highly hypothetical and unrealistic, but our goal here is to have a clear categorization of variables. Specifically, we generated the propensity score using $X_1, \dots, X_7$ and generate the continuous potential outcomes using $X_1, \dots, X_4$ and $X_8, \dots, X_{10}$. Figure 2 illustrates this generating process. As a result, four covariates $(X_1, \dots, X_4)$ were confounders, three $(X_5, \dots, X_7)$ were IVs, three $(X_8, \dots, X_{10})$ were predictors of outcome only, and six $(X_{11}, \dots, X_{16})$ were noise variables. The generating mechanism follows the work of Leacy and Stuart,[11] but with noise variables included.

Figure 3 shows the distributions of the propensity score (on the original scale and logit scale) and prognostic score for both the treatment group and control group. Interestingly, the prognostic score had much better overlap than the propensity score. This was not caused by the special design of the simulation but by the definition and nature of the two scores. This difference in overlap will become an important explanation when we later compare the performance of the three matching estimators in Section 5. It is also worth noting that the difference in overlap relies on the correlation between the propensity score and the prognostic score. In this simple simulation, the correlation is very weak

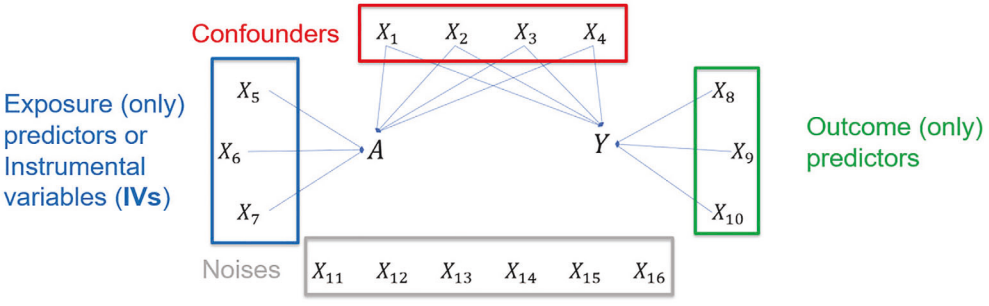


FIGURE 2 Data generation structure of the illustrating example

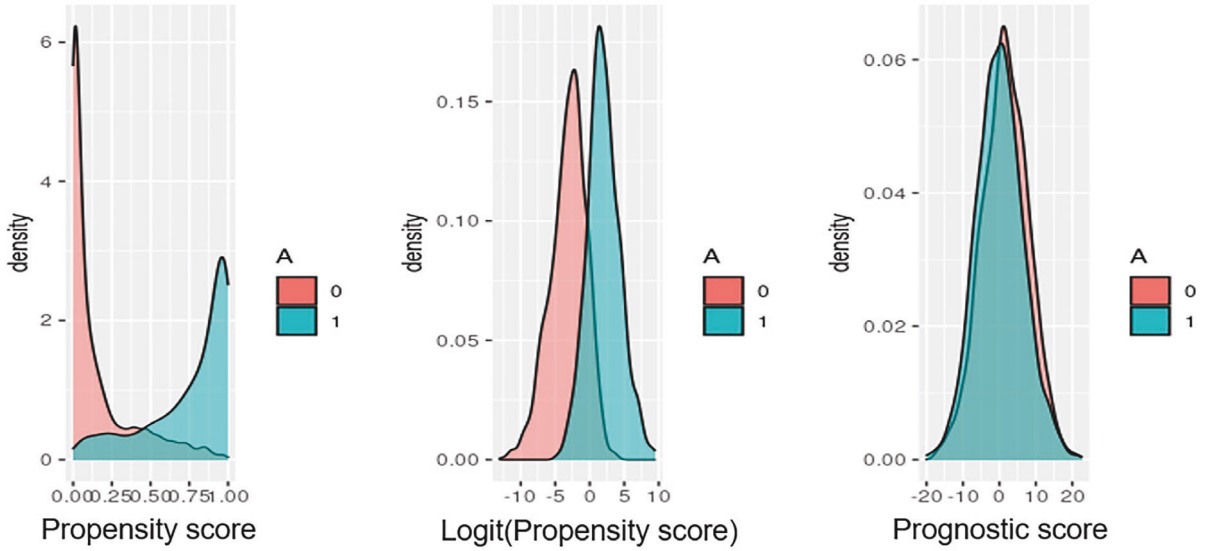


FIGURE 3 Overlapping of propensity score and prognostic score

(around −0.14). When the propensity score is highly correlated with the prognostic score, the difference between PSM and PGM can be negligible.[8] We provide an illustrating example in the supplementary material.

In this simple setting, the individual treatment effect was a constant 3 for all the 3000 subjects in the population. Thus, the ATT was also 3, and we would use PSM, PGM, and DSM to estimate the ATT. For simplicity, we applied one-to-one matching without a caliper and with replacement in all the three methods. In Section 3.2, we will see that the conclusions could be extended to other configurations of caliper and replacement.

Based on different categories of variables, we used the following five sets of covariates to fit each score.

- All variables: $X_1, \ldots, X_{16}$
- All except noise variables: $X_1, \ldots, X_{10}$
- Confounders + IVs: $X_1, \ldots, X_7$
- Confounders + outcome predictors: $X_1, \ldots, X_4$ and $X_8, \ldots, X_{10}$
- Only confounders: $X_1, \ldots, X_4$

Note that we always included confounders when fitting scores, otherwise there would be significant confounding bias. We replicated the simulation 100 times and recorded the estimated ATT from all the three matching methods with all the five variable selection strategies. We seek to find the estimator with the smallest bias and variance. The results will be presented in the following section.

3.1.2 | Results

Figure 4 shows the performance of the PSM estimator under the five variable selection configurations. It can be easily seen that removing instrumental variables is beneficial for PSM: the variance decreased significantly compared to the regular PSM estimator while the bias did not increase much. This is consistent with existing results in the literature.[16,19] Interestingly, even though the propensity score model was misspecified in these two configurations, the estimator was still consistent and even more accurate. This may be because a closer match on IVs is not useful in getting a better match on the potential outcome while overlap of the propensity score becomes poor when including these strong IVs. On the other hand, removing noise variables or outcome predictors does not help reduce the variance but may even increase the bias. This suggests that any model selection method not incorporating outcome may not be effective in improving the PSM estimator. It is also worth mentioning the sample bias observed in the propensity model that includes IVs. Since the theory of PSM guarantees the unbiasedness of the estimator,[3] the observed bias should be a finite sample bias. This was

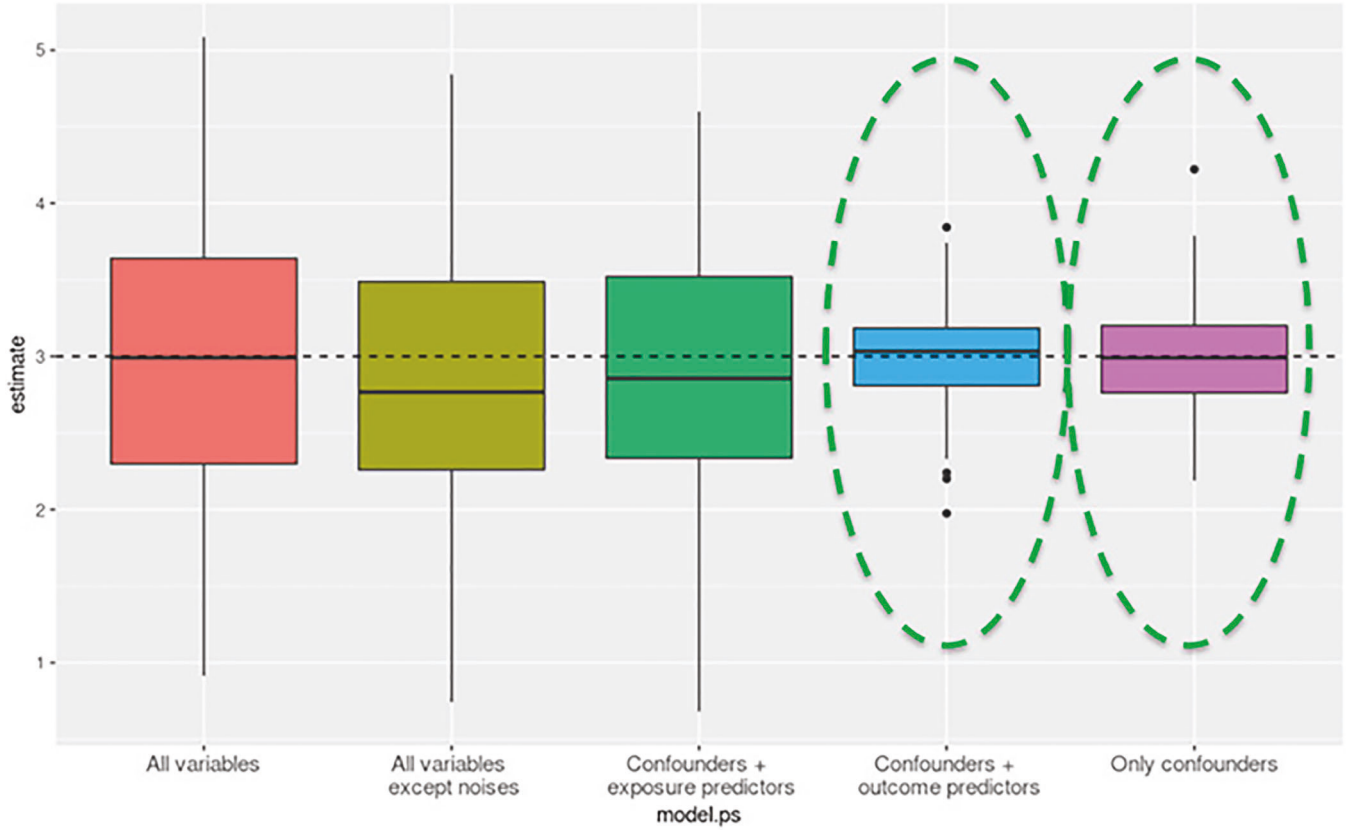


FIGURE 4 Performance of PSM estimator under different variable selection strategies in the illustrating example. Best configurations are marked by green circles

verified by additional simulations reported in the supplementary material. When IVs are included, the finite sample bias may become larger as the variance becomes larger. To sum up, we recommend excluding IVs and noise variables before fitting the propensity score model in PSM.

Figure 5 illustrates the performance of the PGM estimator under the five variable selection strategies. Different from PSM, removing IVs was not very helpful in reducing the variance of the PGM estimator. However, falsely removing outcome predictors could be very harmful: the variance increased a lot, which may increase the finite sample bias of the estimator. This also shows the importance of outcome information. Note that keeping all the variables in the prognostic score model is one of the best configurations. Thus, when fitting the prognostic score model with all covariates is feasible, we recommend skipping the variable selection process in PGM.

In DSM, we have both the propensity score and prognostic score to fit, each with five possible sets of variables. Thus, there are $5 \times 5 = 25$ configurations in DSM. As illustrated in Figure 6, the best strategy is a combination of our previous strategies in PSM and PGM: we should remove IVs from the propensity score model and not remove outcome predictors from the prognostic score model. As a result, we recommend selecting confounders and outcome predictors to fit the propensity score model and using all the variables to fit the prognostic score model.

3.2 | Variable selection in real world: REFLECTIONS dataset

We have figured out the importance and basic strategy of variable selection in matching methods. However, in real datasets, we can never know which variables are outcome predictors and which variables are IVs. As a result, we need some algorithms to select prognostic variables. What's more, various correlation patterns in real-world covariates can increase the complexity of variable categorization. To illustrate these problems, we will introduce the REFLECTIONS dataset and create simulations based on this.

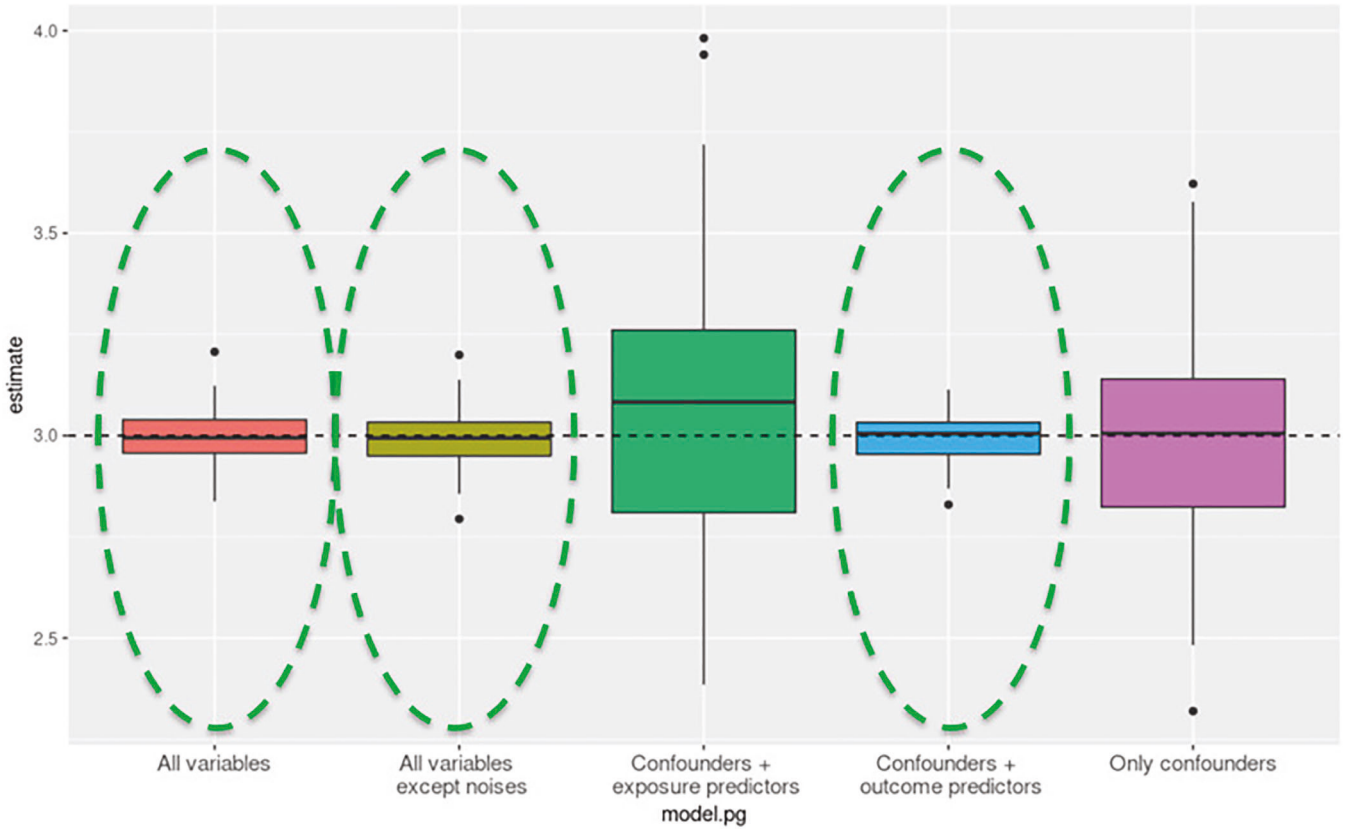


FIGURE 5 Performance of PGM estimator under different variable selection strategies in the illustrating example. Best configurations are marked by green ellipses

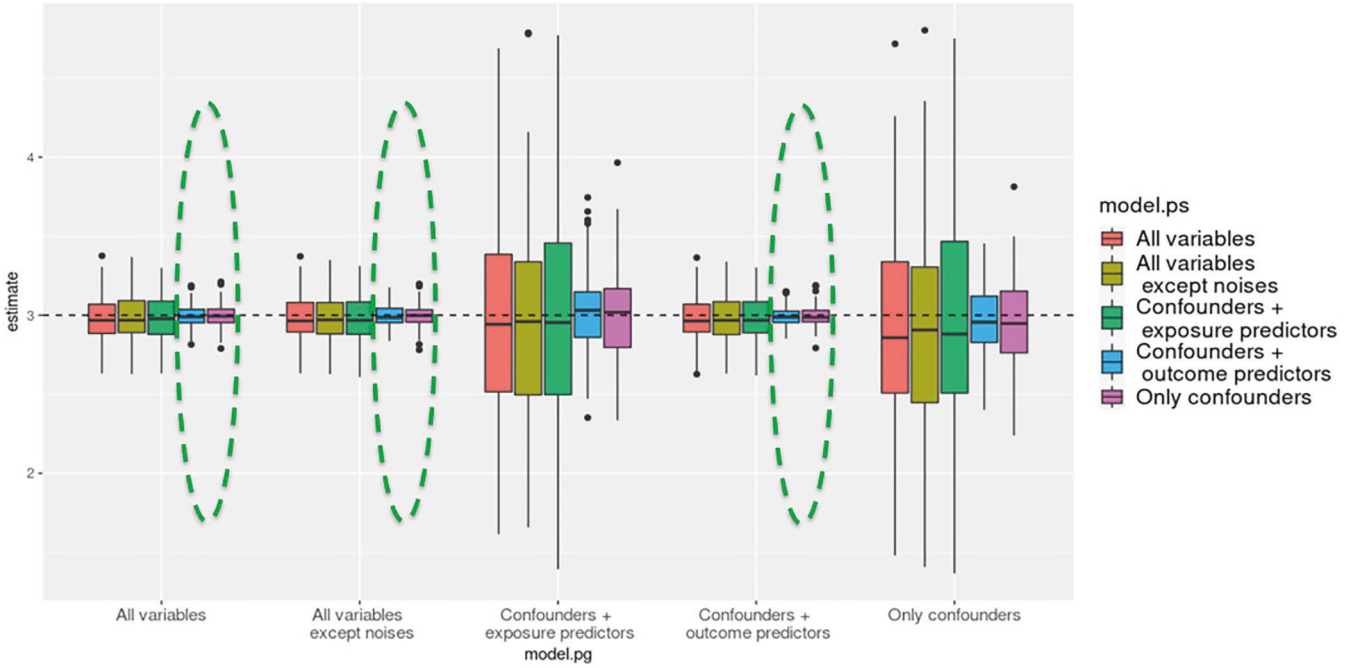


FIGURE 6 Performance of DSM estimator under different variable selection strategies in the illustrating example. Different colors imply different propensity score models, and *x*-axis differentiate various prognostic score models. Best configurations are marked by green ellipses

3.2.1 | Simulated REFLECTIONS dataset

REFLECTIONS stands for the Real World Examination of Fibromyalgia: Longitudinal Evaluation of Cost and Treatments dataset.[30] It was a prospective observational study conducted at 58 clinical sites in the United States and Puerto Rico between 2008 and 2011. The goal was to examine the outcome for patients receiving new treatments for fibromyalgia. The observational data were collected from multiple sources including physician surveys and telephone patient interviews. There were three treatment groups based on patients' treatment at initiation: opioid treatments, non-narcotic opioid like treatment, and all other treatments.[31] In our simulations, we only compare the first two groups so that the treatment is binary. Sixteen continuous variables and eleven categorical variables were recorded for each patient, and we only keep the continuous covariates in our simulations for simplicity. Detailed information for each covariate is exhibited in the supplementary material.

Faries et al[26] implemented the Iman-Conover method to generate simulated REFLECTIONS datasets. Thus, our simulated data did not include any actual data from the REFLECTIONS study. However, the distributions of variables were almost identical in the simulated and real datasets. More importantly, the realistic correlations within the covariates set were retained.

3.2.2 | Methods for variable selection: LASSO

From Section 3.1, we know that it is important to remove IVs and noise variables when fitting the propensity score model, which is equivalent to omitting variables that are irrelevant to the potential outcome. Thus, it is reasonable to apply variable selection algorithms on covariates with respect to the outcome data. For simplicity, here we used LASSO[32] on the control group data, where we consider all the covariates as predictors and the observed outcomes as response. 10-Fold cross-validation was used to select the best tuning parameter λ. We kept two choices of λ: $\lambda_{\min}$ and λ_{1se}, where the first one gave the minimum mean cross-validation error and the latter one was the largest value of λ such that the cross-validation deviance is within 1 standard error of the minimum.[33] After LASSO regression, we would use covariates with nonzero coefficients to fit scores in the matching methods. To sum up, we had three sets of variables to fit the propensity score or the prognostic score:

- All variables
- Variables selected by $\lambda_{\min}$
- Variables selected by λ_{1se}

According to its definition, λ_{1se} is more aggressive in removing variables than $\lambda_{\min}$. As a result, the number of variables used in the selected model decreases when we change from the first strategy to the last strategy. Although more efficient variable selection algorithms could be employed, here we only used LASSO to illustrate the importance of variable selection in the matching methods.

3.2.3 | Simulation setup informed by the REFLECTIONS dataset

We generated 100 different REFLECTIONS datasets based on nonparametric sampling from the original dataset, and there were 3000 individuals in each dataset. Although these datasets were not identical, the correlations among covariates were well retained. Each covariate was standardized before fitting the propensity score and prognostic score. A similar generating structure was used to produce the treatment assignment and potential outcomes, as illustrated in Figure 7. However, because of the complex correlations between covariates, BMI_B, DxDur, PHQ8_B may not be IVs since they may be correlated with other prognostic variables. In fact, every covariate in this REFLECTIONS dataset is a confounder, but some of them can be categorized as IVs or noise variables if they are weakly associated with outcome. Moreover, we opted for two different generating processes for both propensity score and prognostic score: a linear model and a nonlinear model, where the latter case is more difficult with possible overfitting issues in the modeling process. We also considered both constant and heterogeneous treatment effect models in our simulation, where the true effect in the first case was 3. Detailed specifications for the heterogeneous effect as well as models for the propensity

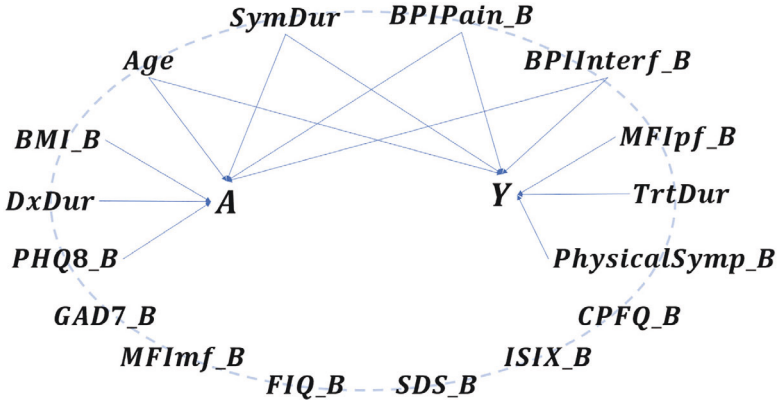


FIGURE 7 Data generation structure of the REFLECTIONS dataset. The dotted ellipse implies complex correlation within the covariates set

and prognostic score can be checked in the supplementary materials. To sum up, there were $2 \times 2 = 4$ scenarios to be considered:

- Linear model + constant effect
- Nonlinear model + constant effect
- Linear model + heterogeneous effect
- Nonlinear model + heterogeneous effect

In each scenario, we used PSM, PGM, and DSM to estimate the ATT. All the possible caliper and replacement choices were considered, and as a result, there were $2 \times 2 = 4$ estimators for each matching method:

- Match with a caliper and with replacement
- Match with a caliper and without replacement
- Match without a caliper and with replacement
- Match without a caliper and without replacement

The caliper was always set as 0.25 times the standard deviation of each score. In DSM, at least one score that satisfies the caliper constraint is sufficient for a matching pair, as explained in Section 2.3. Same as in the illustrating example, we recorded the estimated ATT from each estimator in each scenario, and we sought to find the best configuration with the smallest bias and variance across different situations. Because we were considering the heterogeneous treatment effect, we calculated the bias of all estimates and made the corresponding box-plots, as shown in the following section.

3.2.4 | Results: Variable selection in propensity score matching

Figure 8 presents the results of all the PSM estimators in different scenarios. Clearly, blue boxes were narrower than red and green boxes in most of the scenarios, indicating that variable selection based on the LASSO algorithm with tuning parameter λ_{1se} may significantly increase the accuracy of the estimator. On the other hand, bias may be induced when we remove the weak confounders too aggressively (depending on scenarios). In linear model cases, the bias did not change or even decrease when we used variable selection. This might be because the matching quality was improved when we matched without caliper or the matching rate was increasing when we matched with caliper, see Figure 9. However, when the generating process became a complicated nonlinear model and more higher-order or interaction terms were included, we might falsely remove too many weak confounders and the bias was induced, as illustrated in the nonlinear model with constant treatment effect scenario when we match with replacement. Thus, there was a bias and variance trade-off. We think that variable selection based on λ_{1se} is acceptable because 0 was contained in the blue box and thus

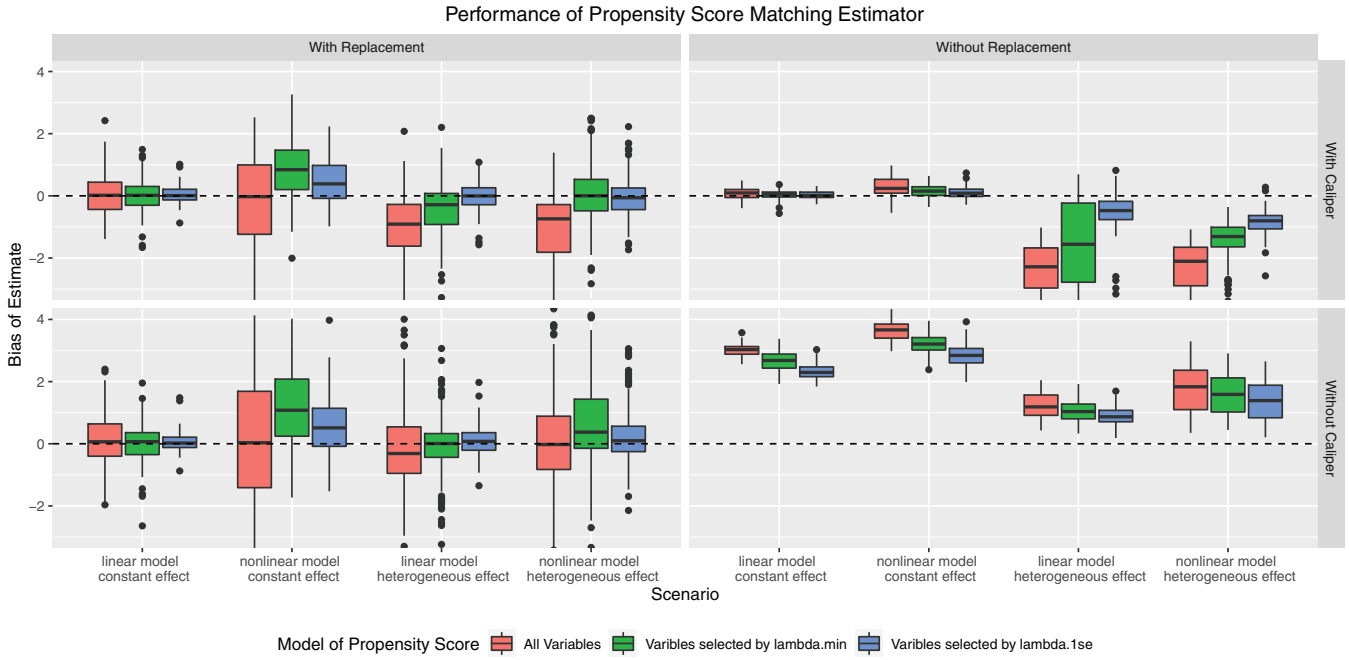


FIGURE 8 Performance of PSM estimator under different variable selection strategies in REFLECTIONS dataset

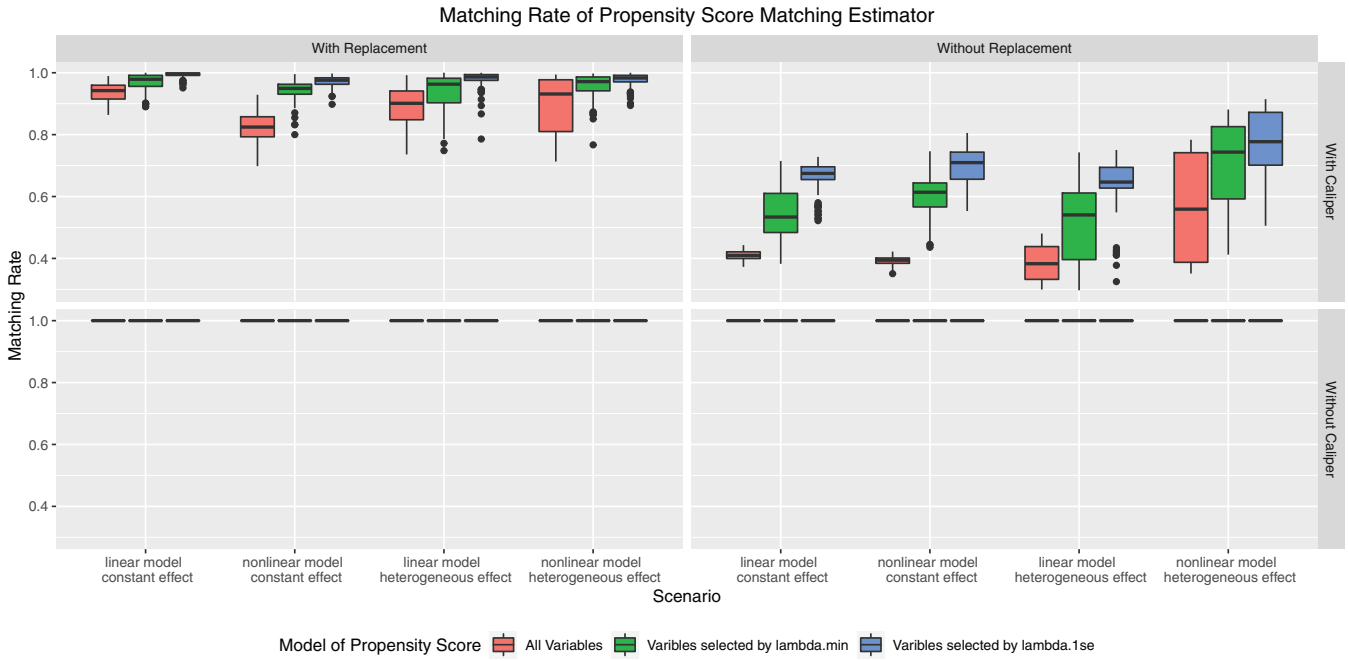


FIGURE 9 Matching rate of PSM estimator under different variable selection strategies in REFLECTIONS dataset

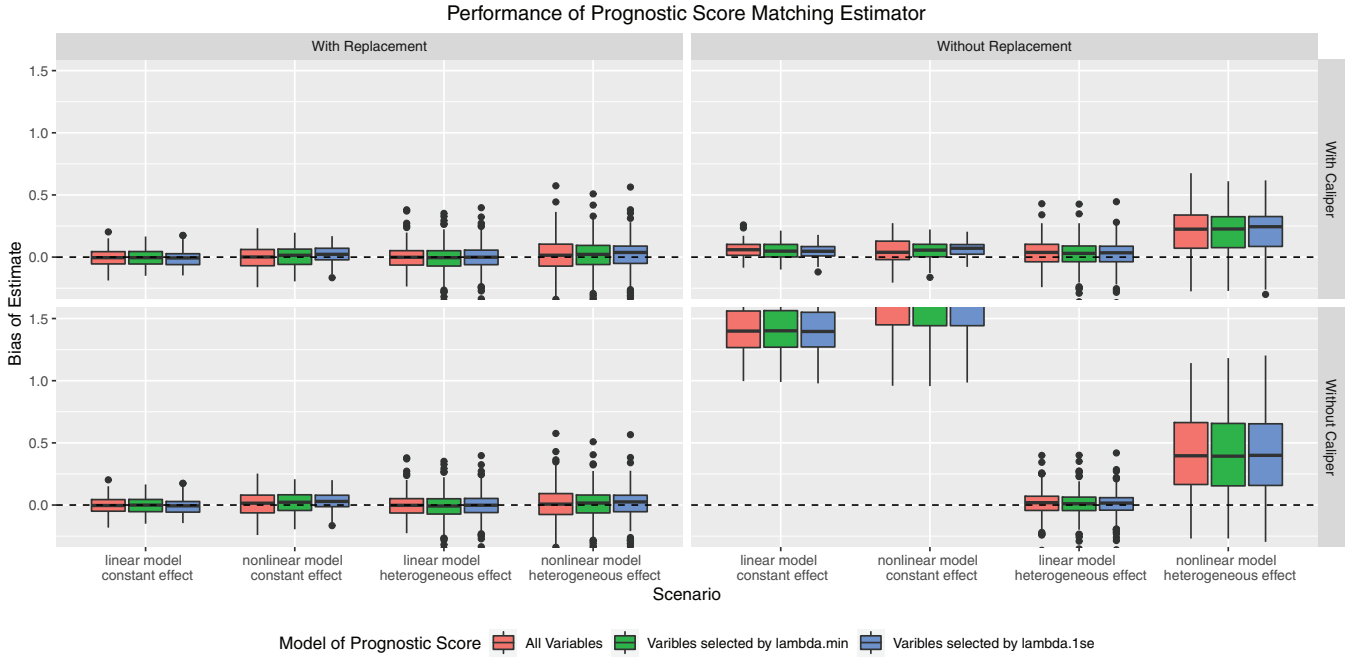


FIGURE 10 Performance of PGM estimator under different variable selection strategies in REFLECTIONS dataset

the bias was not significantly large. It is also worth noting that these were the only two cases when bias increased. In all the other cases the bias decreased or did not change. Thus, we recommend always using LASSO with tuning parameter λ_{1se} to select variables before running PSM.

3.2.5 | Results: Variable selection in prognostic score matching

As illustrated in Figure 10, the performance of PGM was not sensitive to the variable selection strategies. This is consistent with our conclusion in Section 3.1 that there is no significant change in the variance unless we falsely remove outcome predictors. To avoid this risk, we recommend keeping all the variables when we fit the prognostic score in PGM since there is a potential drawback instead of significant improvement from the variable selection.

3.2.6 | Results: Variable selection in double score matching

Similar to Section 3.1, each score could be fit by the three possible sets of variables. In total, there were $3 \times 3 = 9$ variable selection strategies in double score matching. For simplicity, we only showed the results for matching with a caliper and with replacement in the scenario of a linear generating model and heterogeneous treatment effect, see Figure 11. Detailed results are posted in the supplementary materials, and the conclusions are consistent across different scenarios.

Same as the result from Section 3.1.2, variable selection in the propensity score model help reduce the variance, while variable selection in the prognostic score model was not important. From the perspective of bias, it is worth noting that the bias increased as we removed variables from the prognostic score model when we had already selected variables in the propensity score model. The explanation may be that keeping all the variables in the prognostic score model helps avoid the confounding bias induced by removing weak confounders in the propensity score model, which is a special and important property of DSM. Variable selection in PSM is not recommended by some researchers due to the risk of omitting confounders, but DSM can enjoy the benefits from variable selection without taking the risk of confounding bias. As a result, we strongly recommend applying LASSO with tuning parameter λ_{1se} to select variables for the propensity score model but using all the covariates to fit prognostic score in DSM.

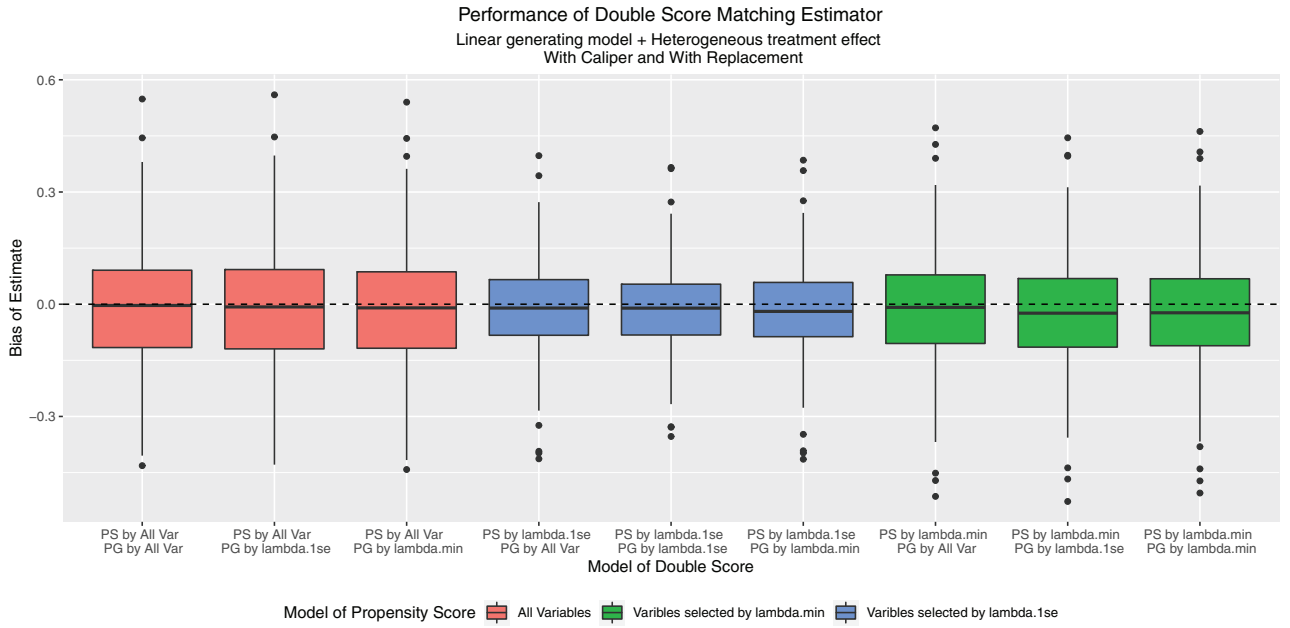


FIGURE 11 Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was linear and the treatment effect was heterogeneous. Only results for matching with a caliper and with replacement are shown here for simplicity

4 | CHOICE OF CALIPER AND REPLACEMENT IN MATCHING

In Section 4, we explored the best variable selection strategy in the matching methods. In this section, we will find the best choice of caliper and replacement based when combined with best model selection strategy from Section 4. That is, we will only compare PSM estimators where the variables are selected by LASSO based on λ_{1se}. The criterion is still that the bias and variance should be small across various situations.

4.1 | Matching constraints in propensity score matching

Figure 12 illustrated the performance of PSM estimators under different caliper and replacement configurations. Matching without replacement reduced the variance of the PSM estimator by removing duplicate samples in the after-match dataset, which was consistent with the findings from the literature.[16,17,19] However, if we matched without a caliper, there was a large bias due to the poor matching quality. If we matched with a caliper, there was still a significant bias when the treatment effect was heterogeneous. This perhaps was because the matched sample from the low matching rate cannot fully represent the target population, and the bias was induced from the difference of estimands. Thus, matching with a caliper and without replacement is only recommended when the constant treatment effect assumption is guaranteed. On the other hand, matching with replacement had a more stable performance because the distance within matching pairs was smaller and the matching rate was higher. Relatively speaking, matching with a caliper could avoid bad matching pairs thus the bias and variance were both smaller. To sum up, we recommend matching with a caliper and with replacement in PSM unless the treatment effect is constant.

4.2 | Matching constraints in prognostic score matching

As shown in Figure 13, matching without replacement was still accompanied by possible bias in PGM. Even in scenarios with constant treatment effect, the bias was still higher compared to matching with replacement. However, in PGM, there is no significant difference between matching with and without a caliper if we match with replacement. Thus, we only need matching with replacement to be guaranteed in PGM.

FIGURE 12 Performance and matching rate of PSM estimator under different choices of caliper and replacement in REFLECTIONS dataset, using variables selected by λ_{1se} from LASSO

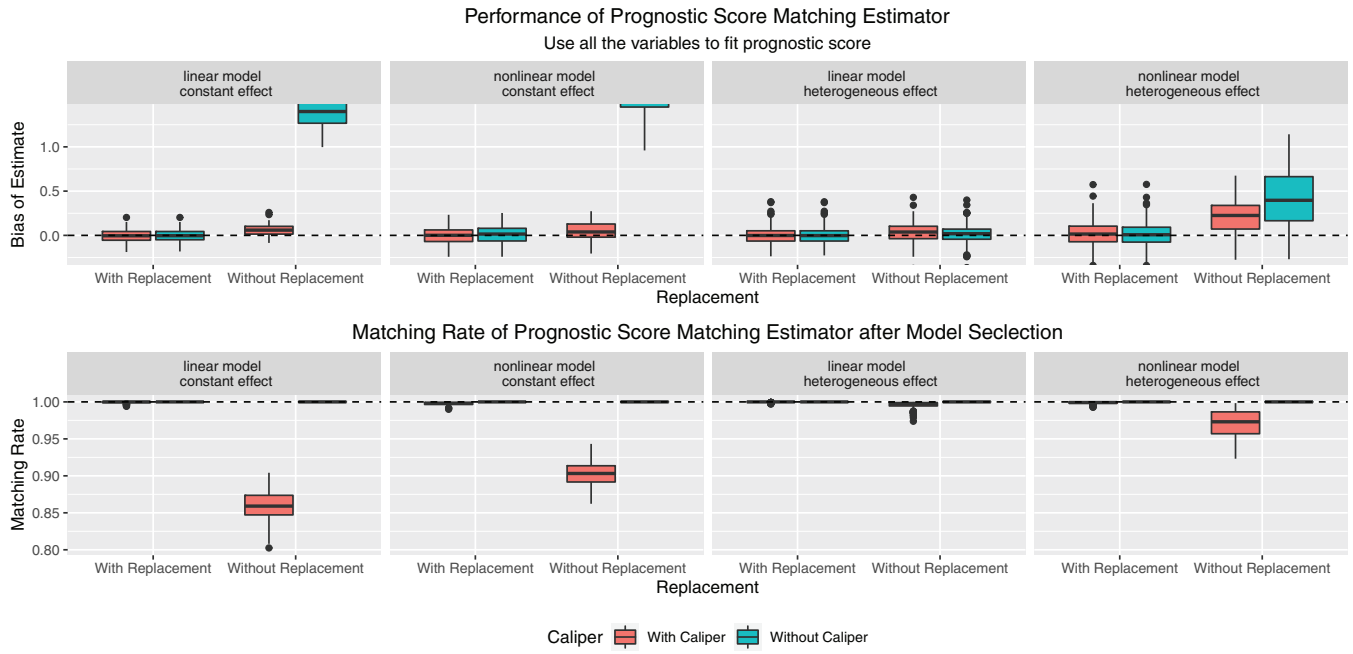


FIGURE 13 Performance and matching rate of PGM estimator under different choices of caliper and replacement in REFLECTIONS dataset, all the variables were used

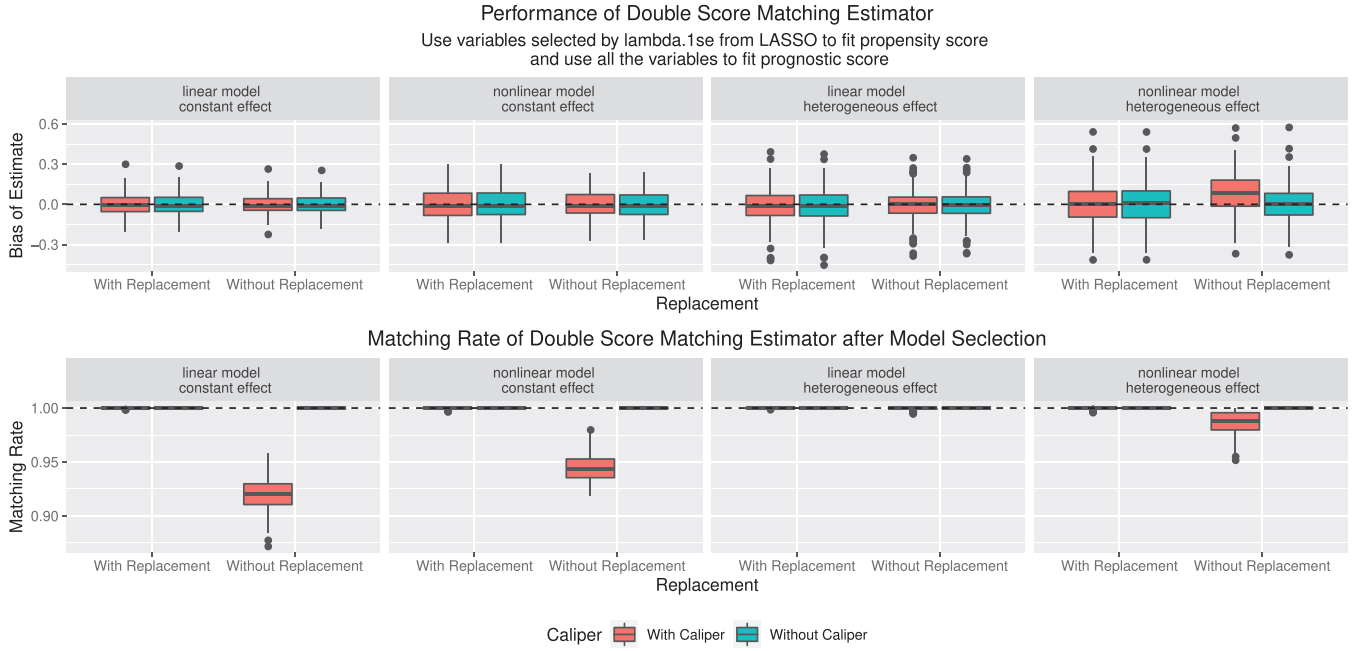


FIGURE 14 Performance and matching rate of DSM estimator under different choices of caliper and replacement in REFLECTIONS dataset, using variables selected by λ_{1se} from LASSO to fit propensity score and all the variables to fit prognostic core

4.3 | Matching constraints in double score matching

Compared to PSM and PGM, DSM was not sensitive to the choice of caliper and replacement, as illustrated in Figure 14. Interestingly, matching without caliper and replacement in DSM had very stable performance across all the scenarios, while PSM and PGM performed extremely badly in this setting. When the generating model was nonlinear and the treatment effect was heterogeneous, adding caliper even increased the bias if we matched without replacement. This is also different from our results in Sections 4.1 and 4.2. Matching without replacement may slightly reduce the variance, while matching with replacement may slightly reduce the bias. However, matching without caliper and replacement may weaken the double robustness of DSM, as shown in the supplementary materials. As a result, we recommend matching with replacement in DSM. Similar to PGM, caliper did not make a significant difference in DSM.

5 | COMPARISON AMONG MATCHING ESTIMATORS

At this point, we have found the best choice of caliper, replacement, and variable selection strategy for PSM, PGM, and DSM. Now we are interested in comparing these three matching methods based on each estimator's best configuration. We will use the same four simulation scenarios as we showed in Section 3.2. However, here we also consider model specification problems. We start with comparing the performance of different matching estimators under correct model specification, and then results for different model misspecification scenarios will be presented in the following section.

5.1 | Comparison under correct model specification

Results under correct model specification are illustrated in Figure 15. In all the four scenarios, PGM and DSM achieved much smaller variance than PSM with negligible bias, suggesting that in practice PSM should be replaced by PGM or DSM if the outcome information is available. Meanwhile, PGM and DSM had very similar performance across different situations. This is different from Leacy and Stuart's result[11] that PGM had a much smaller variance than DSM. The reason for superior performance of DSM in our simulations is that we used the variable selection strategy to improve the accuracy of DSM. The following section will show the advantage of DSM over PGM when models are misspecified.

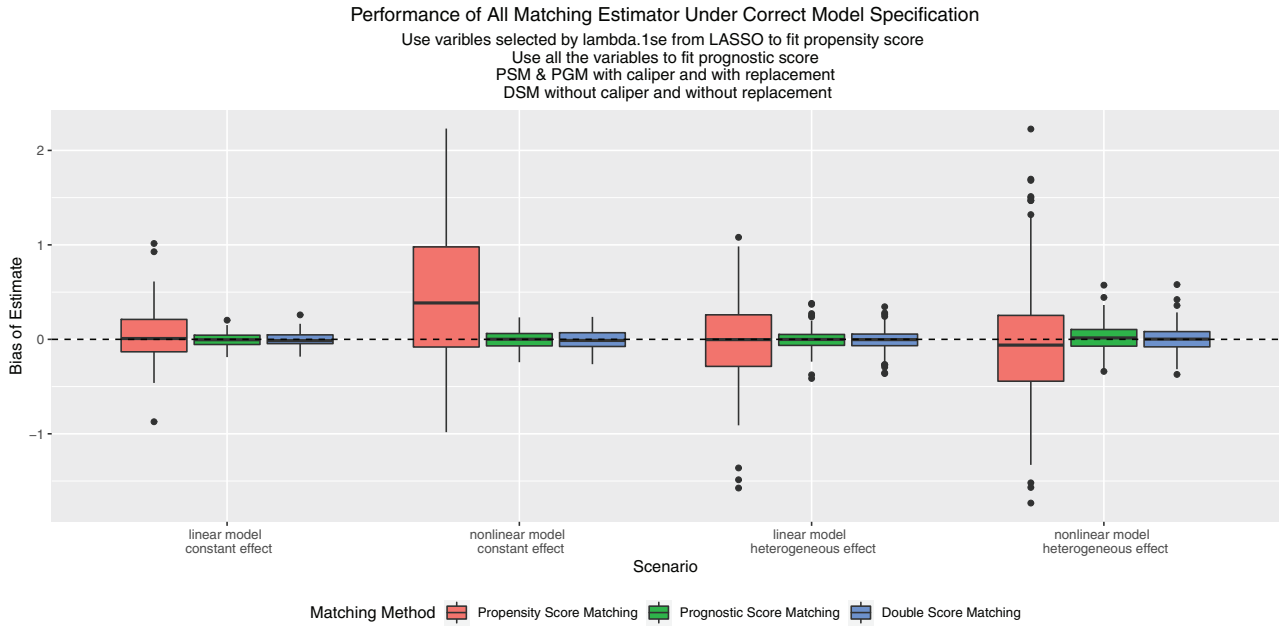


FIGURE 15 Comparison of PSM, PGM, and DSM when both models were correctly specified

5.2 | Comparison under different model specifications

In this section, we consider four different model specifications:

- Both models are correctly specified
- Both models are misspecified
- Only prognostic score model is misspecified
- Only propensity score model is misspecified

In our simulation, model misspecification meant that a linear model was used in the fitting process while the true model was nonlinear. The set of variables included in the fitting model remained the same. Figure 16 presents the performance of the three matching estimators under different model specifications when treatment effect was heterogeneous. PSM and PGM suffered from huge bias if the related score was misspecified, while DSM still exhibited strong performance if only one score was misspecified. This is consistent with the double robustness property of DSM established in the literature. However, the estimators considered here take variable selection into consideration, indicating that variable selection strategy does not compromise the double robustness property of DSM.

6 | CONCLUSIONS AND DISCUSSION

In this article, we explore the statistical and numerical properties of PSM, PGM, and DSM via extensive simulation studies. Based on the simulation results, we summarize the findings for each matching estimator with respect to choices of variable selection strategy, caliper, and replacement:

- PSM can be combined with the variable selection strategy of removing IVs from the propensity score model for efficiency consideration; matching can be done with a caliper and with replacement.
- PGM does not necessarily require variable selection in the sense that variable selection has little impact on efficiency, and matching can be done with replacement, while whether or not using caliper is immaterial.
- DSM can be combined with the variable selection strategy of removing IVs from the propensity score model; matching can be done with replacement, while whether or not using caliper is inconsequential.

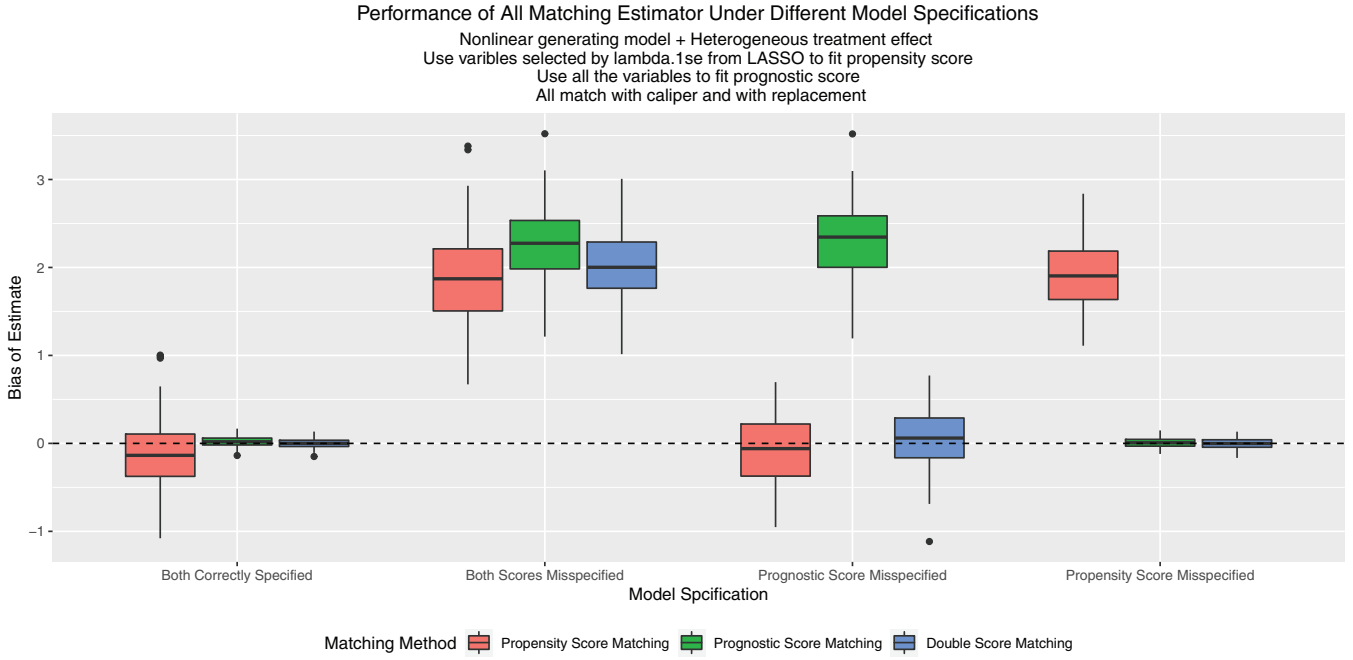


FIGURE 16 Comparison of PSM, PGM, and DSM when both models might be misspecified

Among the three matching schemes, DSM enjoys a double robustness property in that its consistency requires either the propensity score model or the prognostic score model to be correctly specified, in stark contrast to PSM and PGM, indicating that DSM is less vulnerable to model misspecification. Meanwhile, Section 5.1 showed that DSM was the most efficient estimator when models were correctly specified. As a result, DSM is the most robust and efficient matching estimator across different situations. Again, these findings may be restrictive to the simulation settings in the article, where outcomes and covariates are generated from multivariate normal distributions, therefore our recommendations may not apply to other scenarios.

Intrinsically, DSM and PGM require outcome information to fit the prognostic score model, which violates Rubin's principle of designing studies without outcome.[34] The same issue arises when the variable selection strategy is incorporated to remove IVs in PSM. The separation of the design and analysis stages helps establish the credibility of the studies by avoiding the potential for selecting the balancing score based on the outcome results the researcher desires. However, our study showed that when outcome information is available or in retrospective studies, using the outcome information in matching can boost efficiency. Moreover, only the control arm outcome is used when estimating the ATT, and thus outcome for the treatment of interest can remain unobserved. This also safeguards the separation principle and prevents potential data dredging. In prospective studies, prior information from previous research can be used to construct the prognostic score when outcome information is not available.[28]

The variable selection result for PSM is consistent with the finding from Brookhart et al[16] and Myers et al[19] suggesting that including IVs into the propensity score model may amplify bias and variance of the causal effect estimator. However, different from the suggestion from Myers et al[19] that variable selection should not be used in consideration of confounding bias, we still recommend using the variable selection procedure to increase accuracy but using DSM instead of PSM. The prognostic score works as a protection against confounding bias in DSM since all the variables are included in the prognostic score model, implying another important advantage of DSM. In other words, DSM changes a trade-off problem into a win-win situation. However, this may to some extent compromise the double robustness of the DSM estimator when prognostic variables are difficult to identify, such as in studies involving rare outcomes and high-dimensional covariates. It is likely that the variable selection algorithm may remove some important variables improperly, leading DSM to lose its double robustness. From this perspective, there is still a trade-off between efficiency and robustness. We plan to address this issue in our future work. Moreover, although we used LASSO with tuning parameter λ_{1se} to select variables in PSM, we believe that there should be better ways to remove IVs, such as the machine learning algorithms like random forest and neural network. Besides, soft variable selection strategies may also be used in matching methods. For example, Tang et al[35] used causal ball screening to assign higher weights to variables that are more related to outcome. Although this work

was based on the weighting framework, the idea may be applied to matching estimators as well. Since all the variables are kept in the model, it is likely that the double robustness is always guaranteed even in difficult situations such as high-dimensional covariates and rare outcomes. The ball-covariance approach proposed by Pan et al[36] was recently used by Zhao and Yang[37] to select outcome predictors in the matching context. This may also be utilized in our setting. We hope to see more developed variable selection algorithms designated for the matching methods.

We suggest using all the variables in the prognostic score model in PGM and DSM. This conclusion is based on a simulation study with 10 to 100 predictors and 3000 observations. In some extreme scenarios where rare outcomes and high-dimensional covariates are involved, it is necessary to apply some variable selection algorithms since it is impossible to include all the variables in the prognostic score model. The performance may be improved by using more advanced variable selection methods than LASSO, which will be investigated in our future work.

The discussion of caliper was very limited in this article: we only compared caliper of 0.25 standard deviation versus no caliper. Note that the latter case could be seen as a caliper with infinite width. As a result, we were simply considering two special values of caliper width. A more comprehensive study on the choice of caliper can be done in the future. Austin and Wang et al's studies[22,23] are good examples, but a proper variable selection strategy should be applied before matching. Its effect on PGM and DSM should also be studied thoroughly. As far as we observed in our simulations, PGM and DSM were not sensitive to the choice of caliper, and we are interested that whether this will hold for a larger range of caliper widths. For the choice of replacement, the bias observed in matching without replacement under heterogeneous treatment effects may come from the estimation of ATT. When we change the target population from the treatment group to the matched population defined by the matching weights or overlap weights based on the propensity score, matching without replacement may not be biased anymore.[38-40] We also acknowledge that the dichotomy in the article oversimplified the comparison of the existing matching methods. Besides matching with and without replacement that are commonly used, algorithms such as full matching and full matching with structural restrictions are available in some specialized software.[41] In light of this, researchers may choose from a wider range of matching methods. More work can be done in the future to provide practical recommendations for a wider spectrum of matching methods.

Although our final recommendation is DSM, PSM can be still useful when the outcome information is unavailable. In contrast to the usual recommendation that matching should be done with a caliper and without replacement to increase accuracy, we suggest matching with replacement to reduce potential bias from heterogeneous treatment effects. This is a bias-and-variance trade-off, but our simulation showed that the negative impact of bias can be more significant than the problem of variance. Using variable selection strategies may compensate for the loss of accuracy. Despite the fact that the propensity score model is misspecified after removing IVs, the estimator remains consistent. This can be explained by the covariate scores proposed by Waernbaum.[42] What matters is the conditional independence between the potential outcomes and the treatment assignments, which can still hold despite the lack of balance in some covariates across treatment arms.

Our study is limited to the estimation of ATT. When estimating the ATE, multiple prognostic scores are necessary for PGM and DSM. Similar to Corollary 2, Yang and Zhang[14] compared the three estimators for the ATE in their supplementary materials, but they only proved that DSM can be more efficient than PSM when estimating marginal means instead of the ATE. Future work will investigate the performance when variable selection is considered in estimating treatment effects via DSM. However, the definition of IVs may be tricky here. For example, a variable may be irrelevant to the outcome of the control group but important to the outcome of the treated group. As a result, different sets of covariates should be used when estimating each prognostic score. It is not clear how much of a problem this would cause if one just used the prognostic score for the control arm. To deal with this problem, we recommend using the adapted DSM by Yang and Zhang.[14] Instead of directly estimating the ATE, the marginal mean for each treatment level is estimated, where only the propensity score and the prognostic score for that particular level of the treatment need to be adjusted. Thus, we can duplicate the procedure of ATT estimation considered in this article, and different sets of IVs are excluded when estimating the averages of the potential outcomes for the treatment group and control group. We expect to see both theoretical and numerical results on this problem, including the effect of variable selection and the comparison among the three matching estimators. Another limitation is that we were not changing the overlap in propensity score or prognostic score in order to see how it is affecting the DSM performance. But it is worth noting that the prognostic score has good overlap insensitive to model specifications. Moreover, our simulation study is limited to continuous outcomes. Binary outcomes, as well as survival data, can be investigated similarly in the future. Furthermore, the presence of nominal covariates, especially with rare levels, may also change the conclusions in this article since our simulations are based on continuous baseline covariates. These kinds of problems are worth investigating in future research. Researchers should be aware of these limitations and should be cautious when applying our recommendations to their settings. Nevertheless, we believe that our

simulation study based on REFLECTIONS dataset, where variables come from non-normal distributions and exhibited complex correlation structures, is an improvement compared to the existing simulation studies that rely on independent normally distributed variables.

ACKNOWLEDGEMENTS

This research is supported by the National Science Foundation grant DMS 1811245, National Institute of Aging grant 1R01AG066883, and National Institute of Environmental Health Sciences grant 1R01ES031651.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Yunshu Zhang https://orcid.org/0000-0002-2515-2145
Shu Yang https://orcid.org/0000-0001-7703-707X
Douglas E. Faries https://orcid.org/0000-0001-8952-7738
Ilya Lipkovich https://orcid.org/0000-0002-1936-2197

REFERENCES

1. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhyā Ind J Stat Ser A*. 1973;35(4):417-446.
2. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
4. King G, Nielsen RA. Why propensity scores should not be used for matching. *Polit Anal*. 2019;27(4):435-454.
5. Wyss R, Glynn RJ, Gagne JJ. A review of disease risk scores and their application in pharmacoepidemiology. *Current Epidemiol Rep*. 2016;3(4):277-284.
6. Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008;95(2):481-488.
7. Wyss R, Ellis AR, Brookhart MA, et al. Matching on the disease risk score in comparative effectiveness research of new treatments. *Pharmacoepidemiol Drug Saf*. 2015;24(9):951-961.
8. Aikens RC, Greaves D, Baiocchi M. A pilot design for observational studies: using abundant data thoughtfully. *Stat Med*. 2020;39(30):4821-4840.
9. Stuart EA, Lee BK, Leacy FP. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol*. 2013;66(8):S84-S90.
10. Nguyen TL, Debray TP. The use of prognostic scores for causal inference with general treatment regimes. *Stat Med*. 2019;38(11):2013-2029.
11. Leacy FP, Stuart EA. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat Med*. 2014;33(20):3488-3508.
12. Hansen BB. Bias reduction in observational studies via prognosis scores. Technical report 441, University of Michigan, Statistics Department; 2006.
13. Antonelli J, Cefalu M, Palmer N, Agniel D. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*. 2018;74(4):1171-1179.
14. Yang S, Zhang Y. Multiply robust matching estimators of average and quantile treatment effects; 2020. arXiv preprint arXiv:2001.06049.
15. Hu Z, Follmann DA, Qin J. Semiparametric double balancing score estimation for incomplete data with ignorable missingness. *J Am Stat Assoc*. 2012;107(497):247-257.
16. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.
17. Yang S, Kim JK, Song R. Doubly robust inference when combining probability and non-probability samples with high dimensional data. *J Royal Stat Soc Ser B (Stat Methodol)*. 2020;82(2):445-465.
18. De Luna X, Waernbaum I, Richardson TS. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*. 2011;98(4):861-875.
19. Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. 2011;174(11):1213-1222.
20. Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol*. 2011;174(11):1223-1227.
21. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39(1):33-38.
22. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10(2):150-161.
23. Wang Y, Cai H, Li C, et al. Optimal caliper width for propensity score matching of three treatment groups: a Monte Carlo study. *PLoS One*. 2013;8(12):e81045.
24. Imbens GW, Rubin DB. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge University Press; 2015.

25. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J Am Stat Assoc*. 1999;94(448):1053-1062.
26. Faries D, Zhang X, Kadziola Z, et al. *Real World Health Care Data Analysis: Causal Methods and Implementation Using SAS*. Cary, NC: SAS Institute; 2020.
27. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996;52(1):249-264.
28. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcome Res Methodol*. 2001;2(3):169-188.
29. Abadie A, Imbens GW. Matching on the estimated propensity score. *Econometrica*. 2016;84(2):781-807.
30. Robinson RL, Kroenke K, Mease P, et al. Burden of illness and treatment patterns for patients with fibromyalgia. *Pain Med*. 2012;13(10):1366-1376.
31. Peng X, Robinson RL, Mease P, et al. Long-term evaluation of opioid treatment in fibromyalgia. *Clin J Pain*. 2015;31(1):7-13.
32. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
33. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.
34. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20-36.
35. Tang D, Kong D, Pan W, Wang L. Outcome model free causal inference with ultra-high dimensional covariates; 2020. arXiv preprint arXiv:2007.14190
36. Pan W, Wang X, Zhang H, Zhu H, Zhu J. Ball covariance: a generic measure of dependence in Banach space. *J Am Stat Assoc*. 2020;115(529):307-317.
37. Zhao H, Yang S. Outcome-adjusted balance measure for generalized propensity score model selection; 2021. arXiv preprint arXiv:2107.12487.
38. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*. 2013;9(2):215-234.
39. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2018;113(521):390-400.
40. Matsouaka RA, Zhou Y. A framework for causal inference in the presence of extreme inverse probability weights: the role of overlap weights; 2020. arXiv preprint arXiv:2011.01388.
41. Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. *J Comput Graph Stat*. 2006;15(3):609-627.
42. Waernbaum I. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Stat Med*. 2012;31(15):1572-1581.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhang Y, Yang S, Ye W, Faries DE, Lipkovich I, Kadziola Z. Practical recommendations on double score matching for estimating causal effects. *Statistics in Medicine*. 2021;1-25. doi: 10.1002/sim.9289

APPENDIX. THEORY: ASYMPTOTIC RESULTS FOR MATCHING ESTIMATORS ON THE ATT

In this section, we will derive some asymptotic results for the three matching estimators introduced in Section 2. For simplicity, we only consider matching with replacement and without a caliper. Simulations in the following sections will show that this is the optimal setting for matching estimators. Although variable selection strategies such as LASSO are included in the simulations, we do not consider them at this point because they introduce substantial difficulties in theoretical analysis. Nevertheless, the asymptotic results in this section can still provide some insights for choosing the optimal matching estimators.

We first introduce additional notation that is useful in our theoretical analysis. We posit a working model $e(X;\alpha)$ for the propensity score $e(X)$ and a working model $\Psi(X;\beta)$ for the prognostic score $\Psi(X)$. The double score $S(X;\theta) = (e(X;\alpha), \Psi(X;\beta))$ is the combination of these two scores. We denote $\theta^* = (\alpha^*, \beta^*)$ as the true parameter, that is, $S = S(X;\theta^*) = (e(X;\alpha^*), \Psi(X;\beta^*)) = (e(X), \Psi(X))$. For simplicity of exposition, for a generic variable V, denote

$$\mu_a(V) = E\{Y(a) \mid V\}, \quad \sigma_a^2(V) = \mathbb{V}\{Y(a) \mid V\}, \quad e(V) = E(A \mid V),$$

where $\mu_a(V)$ is the mean function for potential outcome $Y(a)$, $\sigma_a^2(V)$ is the variance function, and $e(V)$ is the propensity score given the variable V.

In this article, we only establish the asymptotic results for matching estimators when θ^* is known, following Abadie and Imbens.[29] Although these can be extended for the estimated parameters,[14] the resulting expressions are too cumbersome to make comparison with variances under the estimated $\hat{\theta}$. To formalize our analysis, we first posit some necessary assumptions required by the asymptotic properties of matching estimators.

Assumption 1. There exist constants c_1 and c_2 such that $0 < c_1 \leq e(X) \leq c_2 < 1$ almost surely.

Assumption 2. $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$.

Assumption 3. For $a = 0, 1, (1)$ the matching variable V has a compact and convex support $\mathcal{V}$, with a continuous density bounded and bounded away from zero: there exist constants C_{1L} and C_{1U} such that $C_{1L} \leq f_1(V)/f_0(V) \leq C_{1U}$, for any $V \in \mathcal{V}$; (2) $\mu_a(V)$ and $\sigma_a^2(V)$ satisfy Lipschitz continuity conditions: there exists a constant C_2 such that $\left|\mu_a(V_i) - \mu_a\left(V_j\right)\right| < C_2 \left\|V_i - V_j\right\|$ for any V_i and V_j. and similarly for $\sigma_a^2(V)$; and (3) there exists $\delta > 0$ such that $E\left\{\left|Y(a)^{2+\delta}\right| \mid V\right\}$ is uniformly bounded for any $V \in \mathcal{V}$; (4) $n^{-1/2}\sum_{i=1}^{n}\left\{A_i - (1-A_i)M^{-1}K_{V,i}\right\}\left\{\hat{\mu}_0(V_i) - \mu_0(V_i)\right\} = o_P(1)$.

Assumption 4. The matching variable V deconfounds the control arm outcome from the treatment assignment: $Y(0) \perp\!\!\!\perp A \mid V$

Assumption 1 is the standard positivity assumption, and Assumption 2 implies that there is no unmeasured confounder. Assumption 3 as a regularity and smoothness condition is considered by Abadie and Imbens[29] for the PSM estimator and Yang and Zhang[14] for the DSM estimator. The last statement implies that the nonparametric estimate for the bias correction term is consistent. Assumption 4 shows that the matching variable V is a balancing score, which is satisfied by the propensity score $e(X)$, the prognostic score $\Psi(X)$, and the double score $S(X)$. Importantly, only one model is needed to be correctly specified in the double score to deconfound the potential outcome from the treatment assignment, contributing to the double robustness of DSM.

In the following theorem, we establish the asymptotic result for the matching estimator $\hat{\tau}_{\text{ATT}}$.

Theorem 1. *Under Assumptions 1 to 4, the matching estimator on the ATT based on matching variable V is asymptotically normal:* $\sqrt{n}\left(\hat{\tau}_{\text{ATT}} - \tau_{\text{ATT}}\right) \to \mathcal{N}\left(0, V_{\tau_{\text{ATT}}}\right)$*, in distribution, as* $n \to \infty$*, where*

$$V_{\tau_{\text{ATT}}} = \frac{1}{p^2}E\left[e(V)\left\{\mu_1(V) - \mu_0(V) - \tau_{\text{ATT}}\right\}^2\right] + \frac{1}{p^2}E\left\{e(V)\sigma_1^2(V)\right\} + \frac{1}{p^2}E\left[\sigma_0^2(V)\left\{\frac{e^2(V)}{1-e(V)} + \frac{1}{M}e(V) + \frac{1}{2M}\frac{e^2(V)}{1-e(V)}\right\}\right]$$

and $p = E\{e(X)\} = E\{e(V)\}$ *is the proportion of treatment population. Specifically, V can be the propensity score e(X), the prognostic score* $\Psi(X)$*, and the double score S(X). Importantly, the double score S(X) only requires one of the two scores to be correctly specified, which implies the double robustness of DSM.*

Here M is the number of controlled subjects matched to each treated subject. In our following simulations, M is always chosen as 1. The proof of this theorem is provided in the supplementary material. By replacing V by $e(X)$, $\Psi(X)$, or $S(X)$, we can derive the asymptotic variances for the PSM, PGM, or DSM estimator, respectively. Therefore, it is natural to compare these three variance terms. Unfortunately, there is no deterministic ordering among the variances of the three matching estimators for the ATT. Instead, we consider the estimation of the average control arm outcome on the treated $\mu_{0,\text{trt}} = E\{Y(0) \mid A = 1\}$. It is worth discussing since $\hat{\tau}_{\text{ATT}} = n_1^{-1}\sum_{u=1}^{n}A_iY_i - \hat{\mu}_{0,\text{trt}}$, and the first term is the same for all the matching estimators. The following corollary establishes its asymptotic distribution.

Corollary 1. *Under Assumptions 1 to 4, the matching estimator on the average control arm outcome on the treated based on matching variable V is asymptotically normal:* $\sqrt{n}\left(\hat{\mu}_{0,\text{trt}} - \mu_{0,\text{trt}}\right) \to N\left(0, V_{\mu_{0,\text{trt}}}\right)$*, where*

$$V_{\mu_{0,\text{trt}}} = \frac{1}{p^2}E\left[e(V)\left\{\mu_0(V) - \mu_{0,\text{trt}}\right\}^2\right] + \frac{1}{p^2}E\left[\sigma_0^2(V)\cdot\left\{\frac{e^2(V)}{1-e(V)} + \frac{1}{M}e(V) + \frac{1}{2M}\frac{e^2(V)}{1-e(V)}\right\}\right]$$

.

In the following theorem, we compare DSM with the other two matching estimators under correct model specifications.

Corollary 2. *When models are correctly specified, PGM is always more efficient than DSM for estimation of* $\mu_{0,\text{trt}} = E\{Y(0)|A=1\}$. *DSM is more efficient than PSM if and only if*

$$E\left([\mu_0\{S(X)\} - \mu_0\{e(X)\}]^2\left\{\frac{e^2(X)}{1-e(X)} + \frac{1}{M}e(X) + \frac{1}{2M}\frac{e^2(X)}{1-e(X)} - 1\right\}\right) \geq 0.$$

The first part of Corollary 2 coincides with Leacy and Staurt[11] that PGM has smaller variance than DSM when both models are correctly specified. This underscores the importance of incorporating variable selection into DSM to increase its efficiency, as shown in the simulations.

Supplementary Material for "Best practices of double score matching for estimating causal effects" by Zhang et al.

The supplementary material contains additional proofs and results for the main paper. Section S1, S2, S3 provide the proof of Theorem 1, Corollary 1, and Corollary 2, respectively. Section S4 explains key variables in the REFLECTIONS dataset. Section S5 specifies the detailed configurations for the data generation process in the simulations. Section S8 shows the complete results for variable selection in DSM, as a complement for Section 3.2.6. Section S9 presents additional results to illustrate the weakness of DSM when matching without a caliper and without replacement.

S1 PROOF OF THEOREM 1

Firstly, it is straightforward that

$$\sqrt{n}\left\{\hat{\tau}_{ATT}\left(\theta^*\right)-\tau_{ATT}\right\}=\frac{n}{n_1}\cdot\frac{n_1}{\sqrt{n}}\left\{\hat{\tau}_{ATT}\left(\theta^*\right)-\tau_{ATT}\right\}.$$

Also, because

$$\frac{n}{n_1}\to\frac{1}{p}$$

in probability, we only need to derive the limiting distribution for the second term, which can be expanded into three terms:

$$\begin{aligned}\frac{n_1}{\sqrt{n}}\left\{\hat{\tau}_{ATT}\left(\theta^*\right)-\tau_{ATT}\right\}=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\hat{\mu}_1\left(V_i\right)-\hat{\mu}_0\left(V_i\right)-\tau_{ATT}\right\}\\&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{Y_i-\hat{\mu}_1\left(V_i\right)\right\}\\&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-A_i\right)M^{-1}K_{V,i}\left\{Y_i-\hat{\mu}_0\left(V_i\right)\right\}.\end{aligned}$$

The nonparametric estimators are assumed to be consistent estimates of the true means, thus

$$\begin{aligned}\frac{n_1}{\sqrt{n}}\left\{\hat{\tau}_{ATT}\left(\theta^*\right)-\tau_{ATT}\right\}=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\mu_1\left(V_i\right)-\mu_0\left(V_i\right)-\tau_{ATT}\right\}\\&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{Y_i-\mu_1\left(V_i\right)\right\}\\&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-A_i\right)M^{-1}K_{V,i}\left\{Y_i-\mu_0\left(V_i\right)\right\}+o_P\left(1\right).\end{aligned}$$

Denote

$$T_{1n}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\mu_1\left(V_i\right)-\mu_0\left(V_i\right)-\tau_{ATT}\right\},\tag{S1}$$

$$T_{2n}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{Y_i-\mu_1\left(V_i\right)\right\},\tag{S2}$$

$$T_{3n}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-A_i\right)M^{-1}K_{V,i}\left\{Y_i-\mu_0\left(V_i\right)\right\}.\tag{S3}$$

We first show these three terms all have zero expectations:

$$\begin{aligned} E\left(T_{1n}\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\left(E\left[A_i\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}\right\}|A_i\right]\right) \\ &= \sqrt{n}pE\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}|A_i = 1\right\} \\ &= \sqrt{n}pE\left[E\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}|A_i = 1, V_i\right\}\right] = 0. \end{aligned}$$

The last line is because

$$\tau_{ATT} = E\left\{Y\left(1\right) - Y\left(0\right)|A = 1\right\} = E\left[E\left\{\mu_1\left(V\right) - \mu_0\left(V\right)|A = 1, V\right\}\right].$$

For the second term,

$$\begin{aligned} E\left(T_{2n}\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\left(E\left[A_i\left\{Y_i - \mu_1\left(V_i\right)\right\}|A_i\right]\right) \\ &= \sqrt{n}pE\left\{Y_i - \mu_1\left(V_i\right)|A_i = 1\right\} = 0. \end{aligned}$$

The last line is simply from the definition of $\mu_1\left(V_i\right)$:

$$E\left\{\mu_1\left(V_i\right)|A_i = 1\right\} = E\left[E\left\{Y_i|A_i = 1, V_i\right\}\right] = E\left\{Y_i|A_i = 1\right\}.$$

For the third term,

$$\begin{aligned} E\left(T_{3n}\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1 - A_i\right) M^{-1} K_{V,i}\left\{Y_i - \mu_0\left(V_i\right)\right\} \\ &= \sqrt{n}\left(1 - p\right) E\left[M^{-1} K_{V,i}\left\{Y_i - \mu_0\left(V_i\right)\right\}|A_i = 0\right] \\ &= \sqrt{n}\left(1 - p\right) E\left(E\left[M^{-1} K_{V,i}\left\{Y_i - \mu_0\left(V_i\right)\right\}|A_i = 0, V_i\right]\right) \\ &= \sqrt{n}\left(1 - p\right) E\left(M^{-1} K_{V,i} E\left[\left\{Y_i - \mu_0\left(V_i\right)\right\}|A_i = 0, V_i\right]\right) = 0. \end{aligned}$$

As a result, we prove that the asymptotic bias of $n^{1/2}\left\{\hat{\tau}_{ATT}\left(\theta^*\right) - \tau_{ATT}\right\}$ is zero.
We show that the covariances of T_{1n}, T_{2n}, T_{3n} are zero:

$$\begin{aligned} \operatorname{cov}\left(T_{1n}, T_{2n}\right) =& \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{cov}\left[A_i\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}\right\}, A_j\left\{Y_j - \mu_1\left(V_j\right)\right\}\right] \\ =& \frac{1}{n} \sum_{i=1}^{n} \operatorname{cov}\left[A_i\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}\right\}, A_i\left\{Y_i - \mu_1\left(V_i\right)\right\}\right] \\ =& \frac{1}{n} \sum_{i=1}^{n} E\left(\operatorname{cov}\left[A_i\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}\right\}, A_i\left\{Y_i - \mu_1\left(V_i\right)\right\}|A_i\right]\right) \\ &+ \frac{1}{n} \sum_{i=1}^{n} \operatorname{cov}\left(E\left[A_i\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}\right\}|A_i\right], E\left[A_i\left\{Y_i - \mu_1\left(V_i\right)\right\}|A_i\right]\right) \\ =& pE\left[\operatorname{cov}\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}, Y_i - \mu_1\left(V_i\right)|A_i = 1\right\}\right] + 0 \\ =& pE\left[\operatorname{cov}\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}, Y_i - \mu_1\left(V_i\right)|A_i = 1, V_i\right\}\right] \\ &+ p\operatorname{cov}\left[E\left\{\mu_1\left(V_i\right) - \mu_0\left(V_i\right) - \tau_{ATT}|A_i = 1, V_i\right\}, E\left\{Y_i - \mu_1\left(V_i\right)|A_i = 1, V_i\right\}\right] \\ =& 0. \end{aligned}$$

The other two covariances are automatically zero by construction. Thus, the asymptotic variance is the summation of the three variance terms:

$$\begin{aligned}\mathbb{V}\left(T_{1n}\right) &= E\left(\left[A_i\left\{\mu_1\left(V_i\right)-\mu_0\left(V_i\right)-\tau_{ATT}\right\}\right]^2\right)\\ &= E\left[A_i\left\{\mu_1\left(V_i\right)-\mu_0\left(V_i\right)-\tau_{ATT}\right\}^2|V_i\right]\\ &= E\left[E\left(A_i|V_i\right)\left\{\mu_1\left(V_i\right)-\mu_0\left(V_i\right)-\tau_{ATT}\right\}^2\right]\\ &= E\left[e\left(V\right)\left\{\mu_1\left(V\right)-\mu_0\left(V\right)-\tau_{ATT}\right\}^2\right],\end{aligned}$$

$$\begin{aligned}\mathbb{V}\left(T_{2n}\right) &= E\left(\left[A_i\left\{Y_i-\mu_1\left(V_i\right)\right\}\right]^2\right)\\ &= E\left[A_i\left\{Y_i-\mu_1\left(V_i\right)\right\}^2|V_i\right]\\ &= E\left[E\left(A_i|V_i\right)\left\{Y_i-\mu_1\left(V_i\right)\right\}^2\right]\\ &= E\left[e\left(V_i\right)\left\{Y_i-\mu_1\left(V_i\right)\right\}^2\right]\\ &= E\left\{e\left(V\right)\sigma_1^2\left(V\right)\right\}.\end{aligned}$$

Following Abdie and Imbens[29], the third term has the limiting variance

$$\mathbb{V}\left(T_{3n}\right) \to E\left[\sigma_0^2\left(V\right)\cdot\left\{\frac{e^2\left(V\right)}{1-e\left(V\right)}+\frac{1}{M}e\left(V\right)+\frac{1}{2M}\frac{e^2\left(V\right)}{1-e\left(V\right)}\right\}\right].$$

Combining the three terms and apply the Slutsky's Theorem, this finishes the proof of Theorem 1.

S2 PROOF OF COROLLARY 1

The matching estimator for the average control arm outcome for the treated can be written as

$$\begin{aligned}\hat{\mu}_{0,trt} &= n_1^{-1}\sum_{i=1}^{n}A_i\left\{M^{-1}\sum_{j\in J_{V,i}}Y_j+\hat{\mu}_0\left(V_i\right)-M^{-1}\sum_{j\in J_{V,i}}\hat{\mu}_0\left(V_j\right)\right\}\\ &= n_1^{-1}\sum_{i=1}^{n}A_i\hat{\mu}_0\left(V_i\right)+\left(1-A_i\right)M^{-1}K_{V,i}\left\{Y_i-\hat{\mu}_0\left(V_i\right)\right\}.\end{aligned}$$

Thus,

$$\sqrt{n}\left\{\hat{\mu}_{0,trt}\left(\theta^*\right)-\mu_{0,trt}\right\} = \frac{n}{n_1}\cdot\frac{n_1}{\sqrt{n}}\left\{\hat{\mu}_{0,trt}\left(\theta^*\right)-\mu_{0,trt}\right\},$$

and we only need to derive the asymptotic distribution of the second term:

$$\begin{aligned}\frac{n_1}{\sqrt{n}}\left\{\hat{\mu}_{0,trt}\left(\theta^*\right)-\mu_{0,trt}\right\} =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\mu_0\left(V_i\right)-\mu_{0,trt}\right\}\\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-A_i\right)M^{-1}K_{V,i}\left\{Y_i-\mu_0\left(V_i\right)\right\}+o_P\left(1\right).\end{aligned}$$

Denote

$$\begin{aligned}T'_{1n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_i\left\{\mu_0\left(V_i\right)-\mu_{0,trt}\right\},\\ T'_{3n} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-A_i\right)M^{-1}K_{V,i}\left\{Y_i-\mu_0\left(V_i\right)\right\} = T_{3n}.\end{aligned}$$

Similarly, the covariance of T'_{1n} and T'_{3n} can be shown as zero. The variance of T'_{1n} is

$$\begin{aligned}\mathbb{V}\left(T'_{1n}\right) &= E\left(\left[A_i\left\{\mu_0\left(V_i\right)-\mu_{0,trt}\right\}\right]^2\right)\\ &= E\left[A_i\left\{\mu_0\left(V_i\right)-\mu_{0,trt}\right\}^2|V_i\right]\\ &= E\left[E\left(A_i|V_i\right)\left\{\mu_0\left(V_i\right)-\mu_{0,trt}\right\}^2\right]\\ &= E\left[e\left(V\right)\left\{\mu_0\left(V\right)-\mu_{0,trt}\right\}^2\right].\end{aligned}$$

The variance of T'_{3n} is same as the variance of T_{3n}:

$$\mathbb{V}\left(T'_{3n}\right) = E\left[\sigma_0^2\left(V\right)\cdot\left\{\frac{e^2\left(V\right)}{1-e\left(V\right)}+\frac{1}{M}e\left(V\right)+\frac{1}{2M}\frac{e^2\left(V\right)}{1-e\left(V\right)}\right\}\right].$$

Combining the two terms and apply the Slutsky's Theorem, we finish the proof of Corollary 1.

S3 PROOF OF COROLLARY 2

By Corollary 1, we obtain the asymptotic variance for the the three matching estimators:

$$\begin{aligned}V_{\mu_{0,trt},PSM} &= \frac{1}{p^2}E\left(e\left(X\right)\left[\mu_0\left\{e\left(X\right)\right\}-\mu_{0,trt}\right]^2\right)+\frac{1}{p^2}E\left[\sigma_0^2\left\{e\left(X\right)\right\}\cdot\left\{\frac{e^2\left(X\right)}{1-e\left(X\right)}+\frac{1}{M}e\left(X\right)+\frac{1}{2M}\frac{e^2\left(X\right)}{1-e\left(X\right)}\right\}\right],\\ V_{\mu_{0,trt},PGM} &= \frac{1}{p^2}E\left[e\left(\Psi\right)\left\{\mu_0\left(\Psi\right)-\mu_{0,trt}\right\}^2\right]+\frac{1}{p^2}E\left[\sigma_0^2\left(\Psi\right)\cdot\left\{\frac{e^2\left(\Psi\right)}{1-e\left(S\right)}+\frac{1}{M}e\left(\Psi\right)+\frac{1}{2M}\frac{e^2\left(\Psi\right)}{1-e\left(\Psi\right)}\right\}\right],\\ V_{\mu_{0,trt},DSM} &= \frac{1}{p^2}E\left[e\left(S\right)\left\{\mu_0\left(S\right)-\mu_{0,trt}\right\}^2\right]+\frac{1}{p^2}E\left[\sigma_0^2\left(S\right)\cdot\left\{\frac{e^2\left(S\right)}{1-e\left(S\right)}+\frac{1}{M}e\left(S\right)+\frac{1}{2M}\frac{e^2\left(S\right)}{1-e\left(S\right)}\right\}\right],\end{aligned}$$

where for simplicity we denote $\Psi=\Psi\left(X\right)$ as the prognostic score and $S=S\left(X\right)=\left(e\left(X\right),\Psi\left(X\right)\right)$ as the double score. By the definition of the prognostic score,

$$\begin{aligned}\mu_0\left(S\right) &= E\left\{Y\left(0\right)|S\right\}=E\left\{Y\left(0\right)|e\left(X\right),\Psi\left(X\right)\right\}=E\left\{Y\left(0\right)|X\right\}=\mu_0\left(X\right)\\ \mu_0\left(\Psi\right) &= E\left\{Y\left(0\right)|\Psi\left(X\right)\right\}=E\left\{Y\left(0\right)|X\right\}=\mu_0\left(X\right).\end{aligned}$$

Thus,

$$\mu_0\left(S\right)=\mu_0\left(\Psi\right)=\mu_0\left(X\right). \tag{S4}$$

Similarly,

$$\sigma_0\left(S\right)=\sigma_0\left(\Psi\right)=\sigma_0\left(X\right). \tag{S5}$$

We first compare the variances of PGM and DSM. By (S4) and (S5), we can show that their first terms are equivalent:

$$\begin{aligned}E\left[e\left(S\right)\left\{\mu_0\left(S\right)-\mu_{0,trt}\right\}^2\right] &= E\left[e\left(S\right)\left\{\mu_0\left(\Psi\right)-\mu_{0,trt}\right\}^2\right]\\ &= E\left(E\left[e\left(S\right)\left\{\mu_0\left(\Psi\right)-\mu_{0,trt}\right\}^2|\Psi\right]\right)\\ &= E\left[\left\{\mu_0\left(\Psi\right)-\mu_{0,trt}\right\}^2E\left\{e\left(S\right)|\Psi\right\}\right]\\ &= E\left[e\left(\Psi\right)\left\{\mu_0\left(\Psi\right)-\mu_{0,trt}\right\}^2\right].\end{aligned}$$

The second term in the asymptotic variance of DSM can be written as

$$\begin{aligned}E\left[\sigma_0^2\left(S\right)\left\{\frac{1}{M}e\left(S\right)+\left(1+\frac{1}{2M}\right)\frac{e^2\left(S\right)}{1-e\left(S\right)}\right\}\right] &= E\left(E\left[\sigma_0^2\left(\Psi\right)\left\{\frac{1}{M}e\left(S\right)+\left(1+\frac{1}{2M}\right)\frac{e^2\left(S\right)}{1-e\left(S\right)}\right\}\right]|\Psi\right)\\ &= E\left(\sigma_0^2\left(\Psi\right)E\left[\left\{\frac{1}{M}e\left(S\right)+\left(1+\frac{1}{2M}\right)\frac{e^2\left(S\right)}{1-e\left(S\right)}\right\}\right]|\Psi\right)\\ &= E\left(\sigma_0^2\left(\Psi\right)\left[\frac{1}{M}e\left(\Psi\right)+\left(1+\frac{1}{2M}\right)E\left\{\frac{e^2\left(S\right)}{1-e\left(S\right)}|\Psi\right\}\right]\right).\end{aligned}$$

By Jensen's inequality, we have

$$E\left\{\frac{e^2(S)}{1-e(S)}|\Psi\right\} \geq \frac{e^2(S|\Psi)}{1-e(S|\Psi)} = \frac{e^2(\Psi)}{1-e(\Psi)},$$

where the last equality is simply implied by the construction of the double score:

$$e(S|\Psi) = E\{e(S)|\Psi\} = E[E\{A|S\}|\Psi] = E(A|\Psi) = e(\Psi).$$

As a result, it follows that $V_{\mu_{0,trt},DSM} \geq V_{\mu_{0,trt},PGM}$.

To compare the variances of PSM and DSM, we decompose

$$Y(0) = \mu_0\{e(X)\} + \varepsilon_{0,S|e(X)} + \varepsilon_0, \tag{S6}$$

where $\varepsilon_{0,S|e(X)}$ and ε_0 have mean zero and satisfy that $\mu_0\{e(X)\}, \varepsilon_{0,S|e(X)}, \varepsilon_0$ are mutually independent. Moreover, $\mu_0(S) = \mu_0\{e(X)\} + \varepsilon_{0,S|e(X)}$. With this decomposition, $\sigma_0^2(S) = E\left(\varepsilon_0^2\right)$ and $\sigma_0^2\{e(X)\} = E\left\{\varepsilon_{0,S|e(X)}^2|e(X)\right\} + E\left(\varepsilon_0^2\right)$.

We first expand the first term in $V_{\mu_{0,trt},DSM}$:

$$\begin{aligned}
E\left[e(S)\left\{\mu_0(S_i) - \mu_{0,trt}\right\}^2\right] =& E\left[e(S)\left\{\mu_0\{e(X)\} + \varepsilon_{0,S|e(X)} - \mu_{0,trt}\right\}^2\right] && \text{(S7)}\\
=& E\left(e(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right) && \text{(S8)}\\
&+ 2E\left(e(S)\varepsilon_{0,S|e(X)}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right) && \text{(S9)}\\
&+ E\left\{\varepsilon_{0,S|e(X)}^2\right\}. && \text{(S10)}
\end{aligned}$$

(S8) can be shown to be equivalent as the first term in $V_{\mu_{0,trt},PSM}$:

$$\begin{aligned}
E\left(e(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right) &= E\left\{E\left(e(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2|e(X)\right)\right\}\\
&= E\left(E\{e(S)|e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right)\\
&= E\left(e(X)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]^2\right).
\end{aligned}$$

(S9) is in fact a zero term:

$$\begin{aligned}
E\left(e(S)\varepsilon_{0,S|e(X)}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right) =& E\left(e(S)\left[\mu_0(S) - \mu_0\{e(X)\}\right]\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
=& E\left(e(S)\mu_0(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&- E\left(e(S)\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
=& E\left\{E\left(e(S)\mu_0(S)\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]|e(X)\right)\right\}\\
&- E\left\{E\left(e(S)\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]|e(X)\right)\right\}\\
=& E\left(E[E\{AY(0)|S\}|e(X)]\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&- E\left(E\{e(S)|e(X)\}\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
=& E\left(E\{AY(0)|e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
&- E\left(e(X)\mu_0\{e(X)\}\left[\mu_0\{e(X)\} - \mu_{0,trt}\right]\right)\\
=& 0.
\end{aligned}$$

Thus, the difference between the first terms of $V_{\mu_{0,trt},DSM}$ and $V_{\mu_{0,trt},PSM}$ is simply $E\left\{\varepsilon_{0,S|e(X)}^2\right\}$. Then, by the decomposition from (S6), the second term in $V_{\mu_{0,trt},PSM}$ can be written as:

$$\begin{aligned}
&E\left[\sigma_0^2\{e(X)\}\cdot\left\{\frac{e^2(X)}{1-e(X)} + \frac{1}{M}e(X) + \frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right]\\
=&E\left(\left[E\left\{\varepsilon_{0,S|e(X)}^2|e(X)\right\} + E\left(\varepsilon_0^2\right)\right]\cdot\left\{\frac{e^2(X)}{1-e(X)} + \frac{1}{M}e(X) + \frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right)\\
=&E\left[\sigma_0^2(S)\cdot\left\{\frac{e^2(X)}{1-e(X)} + \frac{1}{M}e(X) + \frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right]\\
&+ E\left[E\left\{\varepsilon_{0,S|e(X)}^2|e(X)\right\}\left\{\frac{e^2(X)}{1-e(X)} + \frac{1}{M}e(X) + \frac{1}{2M}\frac{e^2(X)}{1-e(X)}\right\}\right].
\end{aligned}$$

As a result, the difference between $V_{\mu_{0,trt},DSM}$ and $V_{\mu_{0,trt},PSM}$ is

$$\begin{aligned} V_{\mu_{0,trt},psm} - V_{\mu_{0,trt},dsm} &= \frac{1}{p^2} E\left[E\left\{ \varepsilon^2_{0,S|e(X)} | e(X) \right\} \left\{ \frac{e^2(X)}{1-e(X)} + \frac{1}{M} e(X) + \frac{1}{2M} \frac{e^2(X)}{1-e(X)} - 1 \right\} \right] \\ &= \frac{1}{p^2} E\left[\varepsilon^2_{0,S|e(X)} \left\{ \frac{e^2(X)}{1-e(X)} + \frac{1}{M} e(X) + \frac{1}{2M} \frac{e^2(X)}{1-e(X)} - 1 \right\} \right] \\ &= \frac{1}{p^2} E\left(\left[\mu_0(S) - \mu_0\{e(X)\} \right]^2 \left\{ \frac{e^2(X)}{1-e(X)} + \frac{1}{M} e(X) + \frac{1}{2M} \frac{e^2(X)}{1-e(X)} - 1 \right\} \right). \end{aligned}$$

Unfortunately, this difference can be either positive or negative. Thus, we don't have a deterministic conclusion on the comparison between PSM and DSM. However, simulations show that DSM can be more efficient than PSM in most situations.

S4 KEY VARIABLES IN THE REFLECTIONS DATASET

Table S1 as an excerpt from Faries et al.'s book[26] summarizes all the continuous variables from the REFLECTIONS dataset that were used in the simulations. Detailed explanation for each variable can be found in the Peng et al. analysis[31].

TABLE S1 List of the continuous variables in the REFLECTIONS dataset.

Variable Name	Variable Label
Age	Age in years
SymDur	BMI at Baseline
BPIPain_B	BPI Pain score at Baseline
BPIInterf_B	BPI Interference score at Baseline
BMI_B	BMI at Baseline
DxDur	Time (in years) since initial Dx
PHQ8_B	PHQ8 total score
MFIpf_B	MFI Physical Fatigue at Baseline
TrtDur	Time (in years) since initial Trtmnt
PhysicalSymp_B	PHQ 15 total score at Baseline
GAD7_B	GAD7 total score at Baseline
CPFQ_B	CPFQ Total Score at Baseline
FIQ_B	FIQ Total Score at Baseline
SDS_B	SDS total score at Baseline
ISIX_B	ISIX total score at Baseline
CPFQ_B	CPFQ Total score at Baseline

S5 DETAILED CONFIGURATIONS FOR THE DATA GENERATION

S5.1 Model specification for the illustrating example

The covariates $X_1 - X_{16}$ are independent standard normal variables.

The true propensity score model is

$$\text{logit}\{P(A=1 \mid \mathbf{X})\} = -1 + 1.5X_1 - 0.5X_2 + 1.75X_3 + 0.5X_4 + 1.5X_5 - 1.25X_6 + 1.75X_7.$$

The true outcome models are

$$Y(0) = 1 - 2.5X_1 - 1.5X_2 - X_3 + 3X_4 + 1.5X_8 - 3.5X_9 + 2.5X_{10} + \epsilon$$
$$Y(1) = Y(0) + 3,$$

where $\epsilon \sim \mathcal{N}(0, 1)$. This error term is also contained in the following section.

S5.2 Model specification for the REFLECTIONS simulations

The covariates are selected from the simulated RELFECTIONS datasets, which is generated from the real REFLECTIONS datasets using the Iman-Conover method[26]. The true propensity score and outcome models vary across different scenarios.

S5.2.1 Linear model and constant effect

The true propensity score is

$$\begin{aligned}\text{logit}\{P(A = 1 \mid \mathbf{X})\} = (&-2 + Age - 0.5SymDur + 1.5BPIPain\_B + 0.5BPIInterf\_B \\ &+ 3BMI\_B - 2.5DxDur + 3.5PHQ8\_B)/2.\end{aligned}$$

The true outcome models are

$$\begin{aligned}Y(0) =& 1 - 2Age - SymDur - 0.5BPIPain\_B + 3BPIInterf\_B \\ &+ 1.5MFIpf\_B - 3.5TrtDur + 2.5PhysicalSymp\_B + \epsilon \\ Y(1) =& Y(0) + 3,\end{aligned}$$

S5.2.2 Nonlinear model and constant effect

The true propensity score is

$$\begin{aligned}\text{logit}\{P(A = 1 \mid \mathbf{X})\} = (&-6 + 3Age - 0.5SymDur + 1.5BPIPain\_B + 0.5BPIInterf\_B + 3BMI\_B \\ &- 2.5DxDur + 3.5PHQ8\_B + Age \cdot BPIPain\_B + 1.5SymDur \cdot BPIInterf\_B \\ &+ 0.5BPIPain\_B \cdot BMI\_B + 1.5BPIInterf\_B \cdot DxDur + 2.5BMI\_B \cdot PHQ8\_B \\ &+ 1.5Age \cdot DxDur + 2SymDur \cdot BPIPain\_B + 0.5BPIPain\_B \cdot BPIInterf\_B \\ &+ BPIInterf\_B \cdot BMI\_B + 1.5BMI\_B \cdot DxDur - 2SymDur \cdot SymDur \\ &+ 1.5BPIInterf\_B \cdot BPIInterf\_B + 1.5DxDur \cdot DxDur)/8.\end{aligned}$$

The true outcome models are

$$\begin{aligned}Y(0) =& 1 - 2Age - SymDur - 0.5BPIPain\_B + 3BPIInterf\_B + 1.5MFIpf\_B - 3.5TrtDur + 2.5PhysicalSymp\_B \\ &+ SymDur \cdot BPIInterf\_B + 2.5BPIPain\_B \cdot PHQ8\_B + BPIInterf\_B \cdot TrtDur \\ &+ 1.5MFIpf\_B \cdot PhysicalSymp\_B + 2Age \cdot TrtDur + 0.5SymDur \cdot BPIPain\_B \\ &+ BPIPain\_B \cdot BPIInterf\_B + 1.5BPIInterf\_B \cdot MFIpf\_B + 0.5MFIpf_B \cdot TrtDur \\ &- SymDur \cdot SymDur + 2.5BPIInterf\_B \cdot BPIInterf\_B + 1.5PhysicalSymp\_B \cdot PhysicalSymp\_B + \epsilon \\ Y(1) =& Y(0) + 3,\end{aligned}$$

S5.2.3 Linear model and heterogeneous effect

The true propensity score is

$$\begin{aligned}\text{logit}\{P(A = 1 \mid \mathbf{X})\} = (&-3 + 6Age - 0.5SymDur + 1.5BPIPain\_B + 0.5BPIInterf\_B \\ &+ 5BMI\_B - 2.5DxDur + 3.5PHQ8\_B)/2.\end{aligned}$$

The true outcome models are

$$\begin{aligned} Y(0) =& 1 - 2Age - SymDur - 0.5BPIPain\_B + 3BPIInterf\_B \\ &+ 1.5MFIpf\_B - 3.5TrtDur + 2.5PhysicalSymp\_B + \epsilon \\ Y(1) =& Y(0) + 3 + 5Age + 4BMI\_B - 3MFIpf\_B - 2.5GAD7\_B. \end{aligned}$$

S5.2.4 Nonlinear model and heterogeneous effect

The true propensity score is

$$\begin{aligned} \text{logit}\{P(A=1 \mid \mathbf{X})\} = (& -6 + 3Age - 0.5SymDur + 1.5BPIPain\_B + 0.5BPIInterf\_B + 3BMI\_B \\ & - 2.5DxDur + 3.5PHQ8\_B + Age \cdot BPIPain\_B + 1.5SymDur \cdot BPIInterf\_B \\ & + 0.5BPIPain\_B \cdot BMI\_B + 1.5BPIInterf\_B \cdot DxDur + 2.5BMI\_B \cdot PHQ8\_B \\ & + 1.5Age \cdot DxDur + 2SymDur \cdot BPIPain\_B + 0.5BPIPain\_B \cdot BPIInterf\_B \\ & + BPIInterf\_B \cdot BMI\_B + 1.5BMI\_B \cdot DxDur - 2SymDur \cdot SymDur \\ & + 1.5BPIInterf\_B \cdot BPIInterf\_B + 1.5DxDur \cdot DxDur)/8. \end{aligned}$$

The true outcome models are

$$\begin{aligned} Y(0) =& 1 - 2Age - SymDur - 0.5BPIPain\_B + 3BPIInterf\_B + 1.5MFIpf\_B - 3.5TrtDur + 2.5PhysicalSymp\_B \\ & + SymDur \cdot BPIInterf\_B + 2.5BPIPain\_B \cdot PHQ8\_B + BPIInterf\_B \cdot TrtDur \\ & + 1.5MFIpf\_B \cdot PhysicalSymp\_B + 2Age \cdot TrtDur + 0.5SymDur \cdot BPIPain\_B \\ & + BPIPain\_B \cdot BPIInterf\_B + 1.5BPIInterf\_B \cdot MFIpf\_B + 0.5MFIpf_B \cdot TrtDur \\ & - SymDur \cdot SymDur + 2.5BPIInterf\_B \cdot BPIInterf\_B + 1.5PhysicalSymp\_B \cdot PhysicalSymp\_B + \epsilon \\ Y(1) =& Y(0) + 3 + 5Age + 4BMI\_B - 2MFIpf\_B - 1.5GAD7\_B \\ & + 2Age \cdot Age + BMI\_B \cdot BMI\_B - 0.5MFIpf\_B \cdot MFIpf\_B \\ & + Age \cdot BMI\_B - MFIpf\_B \cdot GAD7\_B + 1.5Age \cdot GAD7\_B - 2BMI\_B \cdot MFIpf\_B, \end{aligned}$$

S6 PERFORMANCE OF THE MATCHING ESTIMATORS WHEN THE PROPENSITY SCORE AND THE PROGNOSTIC SCORE ARE HIGHLY CORRELATED

In Section 3.1.1, we stated that the prognostic score may not have better a overlap than the propensity score when the two scores are highly correlated. Thus, the difference between PSM and PGM can be negligible. In this section, we construct an extreme example that the prognostic score is exactly the same as the linear predictor of the propensity score using the model specification in Section S5.1. In this case, the overlap of the true prognostic score should be the same as the overlap of the true propensity score in the logit scale. We also check the performance of PSM, PGM, and DSM based on the estimated scores. As illustrated in Figure S1, PSM performed almost equivalently as PGM. DSM had a smaller bias than the other two estimators, but the improvement is not significant. Even though the prognostic score cannot increase the efficiency in this setting, the performance of DSM won't be worse than the two single score estimators.

S7 ADDITIONAL RESULTS FOR VARIABLE SELECTION IN PSM AND PGM

To verify that the bias in the simulation studies in Section 3 is finite sample bias, we increase the number of replications from 100 to 500. The sample bias decreased significantly as illustrated in Figure S2 and S3.

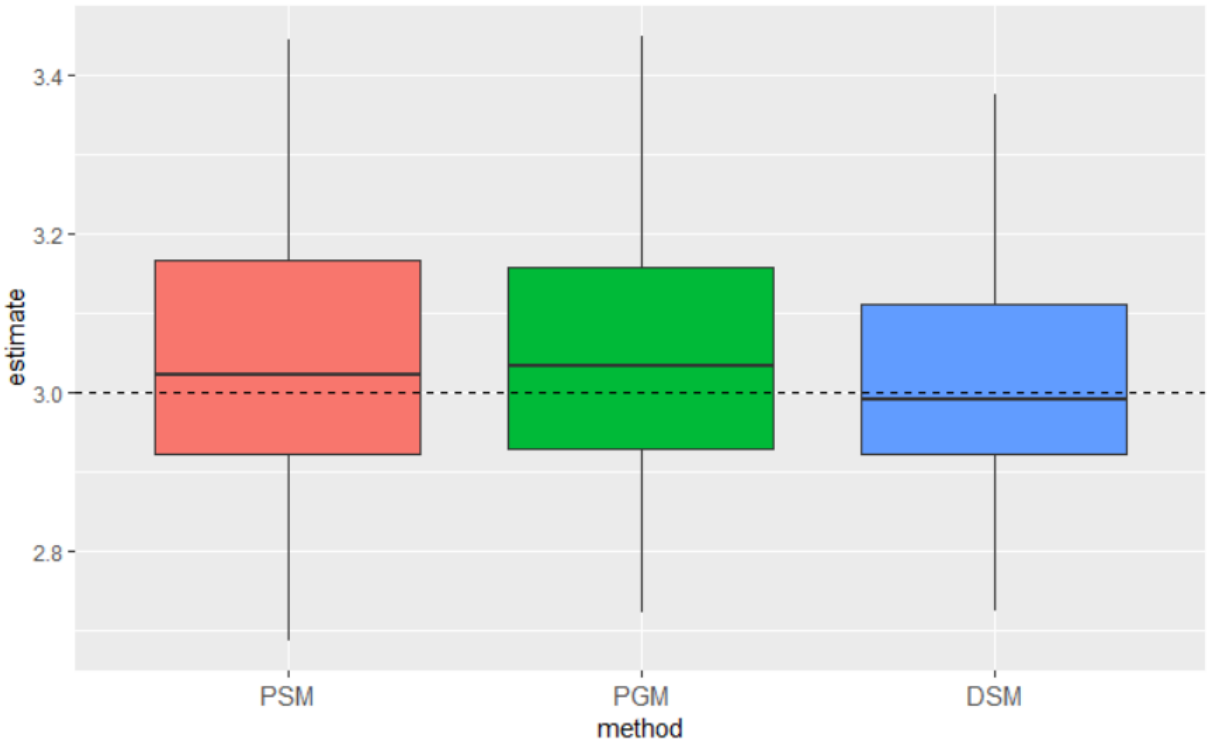


FIGURE S1 Comparison of PSM, PGM, and DSM when the propensity score is equivalent as the prognostic score. Matching is done with caliper and with replacement.

S8 COMPLETE RESULTS FOR VARIABLE SELECTION IN DOUBLE SCORE MATCHING

In Section 4.3, we only showed the results for matching with a caliper and with replacement when the generating model was linear and the treatment effect was heterogeneous. To complete our statements, Figure S4 - S7 present the results for all the scenarios, and our conclusion for DSM holds under different situations.

S9 ADDITIONAL RESULTS FOR DOUBLE SCORE MATCHING WITHOUT A CALIPER AND WITHOUT REPLACEMENT

In Section 4.3, we recommended matching with replacement for DSM, while the caliper did not make a difference. However, as illustrated by Figure 14, matching without a caliper and without replacement achieved the best performance as well. The reason that we excluded this configuration is that the double robustness property of DSM was weakened under this choice of caliper and replacement, as illustrated in Figure S8. When the propensity score was correctly specified but the prognostic score was misspecified, significant bias appeared in DSM. However, this phenomenon did not happen when we match with replacement. As a result, we do not recommend using DSM without a caliper and without replacement, even though it has satisfying performance when models are correctly specified.

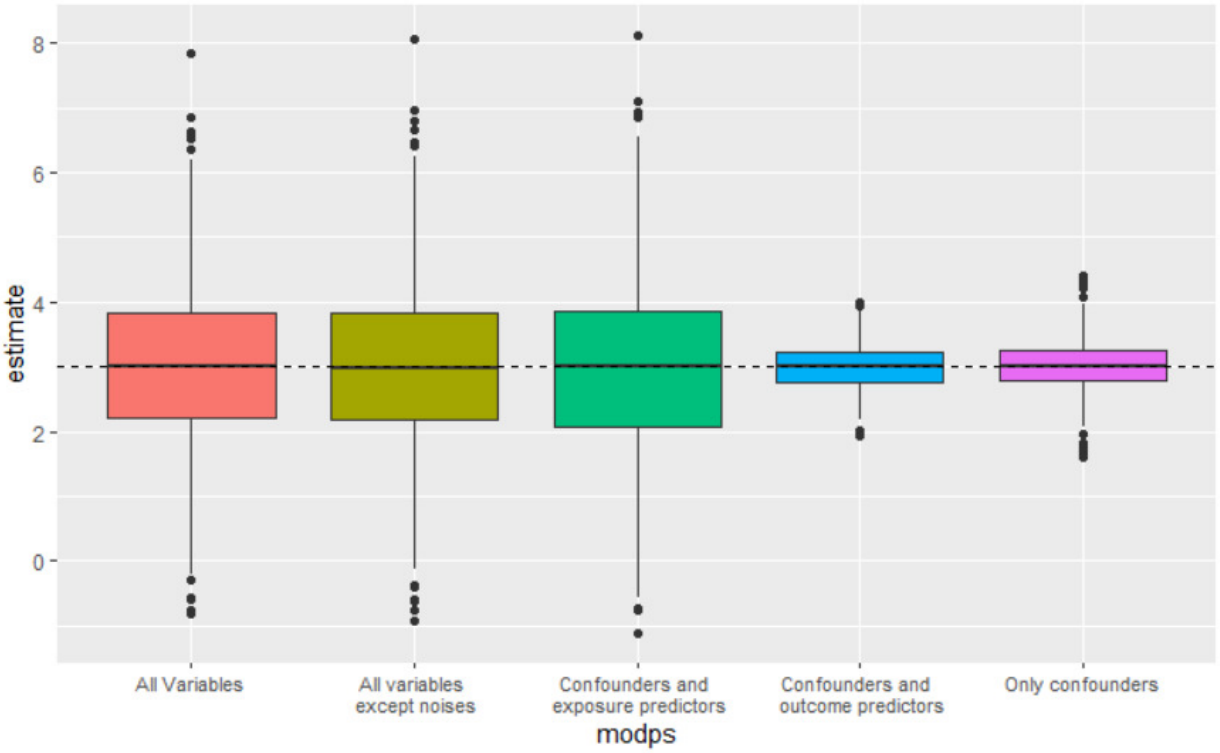


FIGURE S2 Performance of PSM estimator under different variable selection strategies in the illustrating example, number of replications $K = 500$.

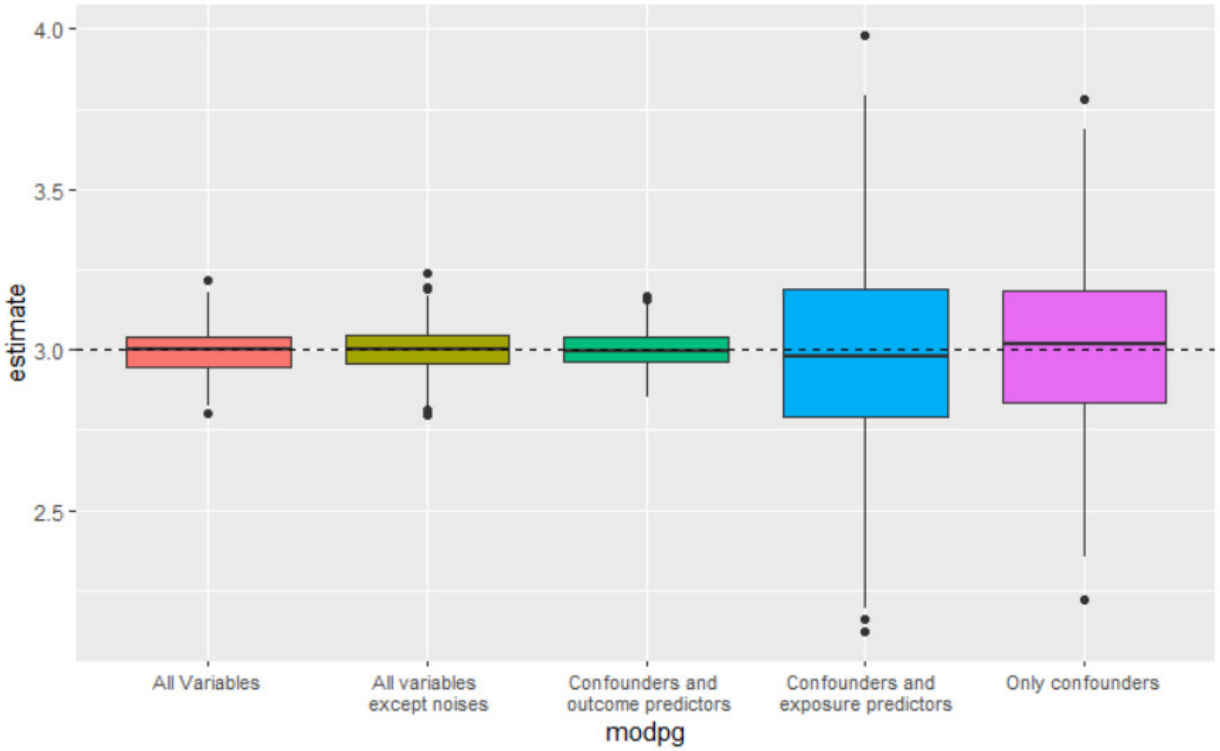


FIGURE S3 Performance of PGM estimator under different variable selection strategies in the illustrating example, number of replications $K = 500$.

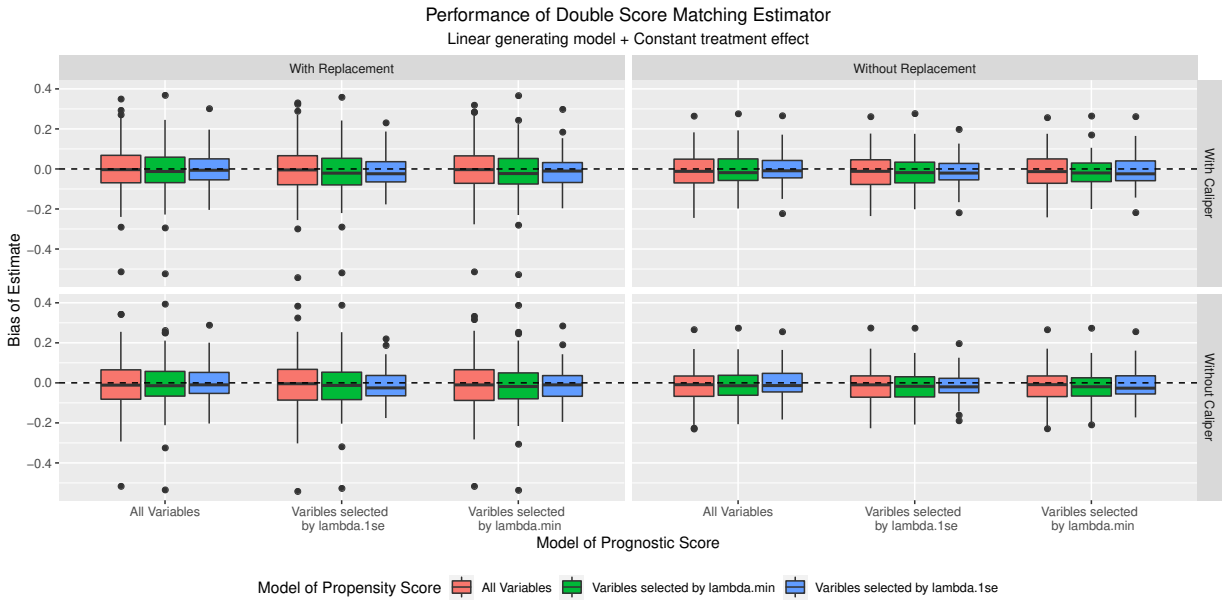


FIGURE S4 Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was linear and the treatment effect was constant.

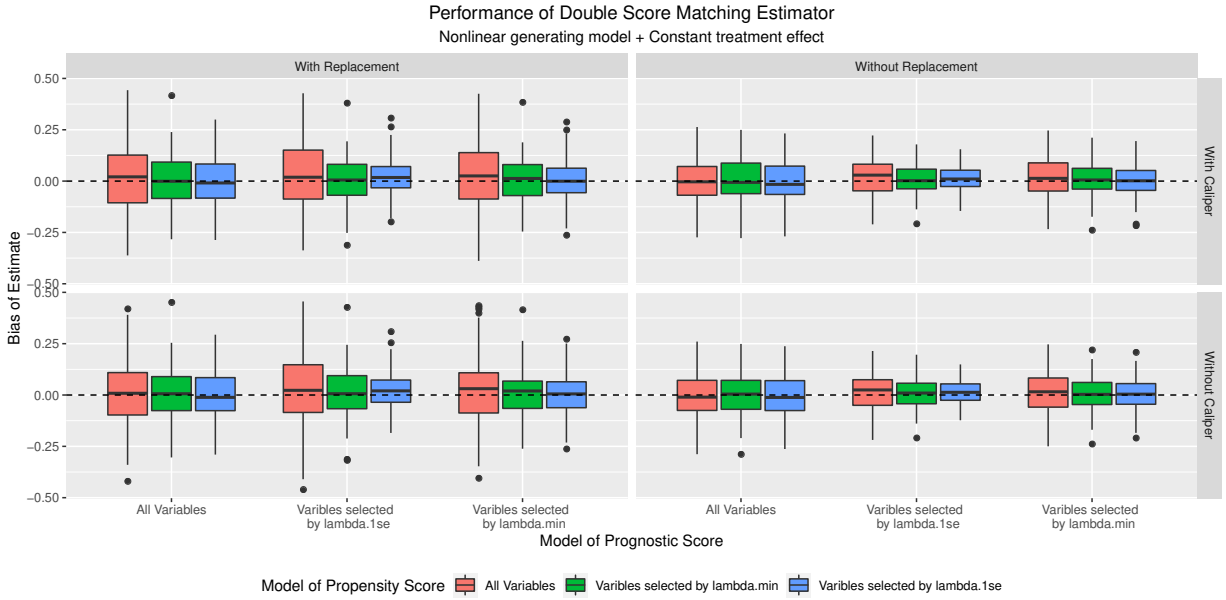


FIGURE S5 Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was nonlinear and the treatment effect was constant.

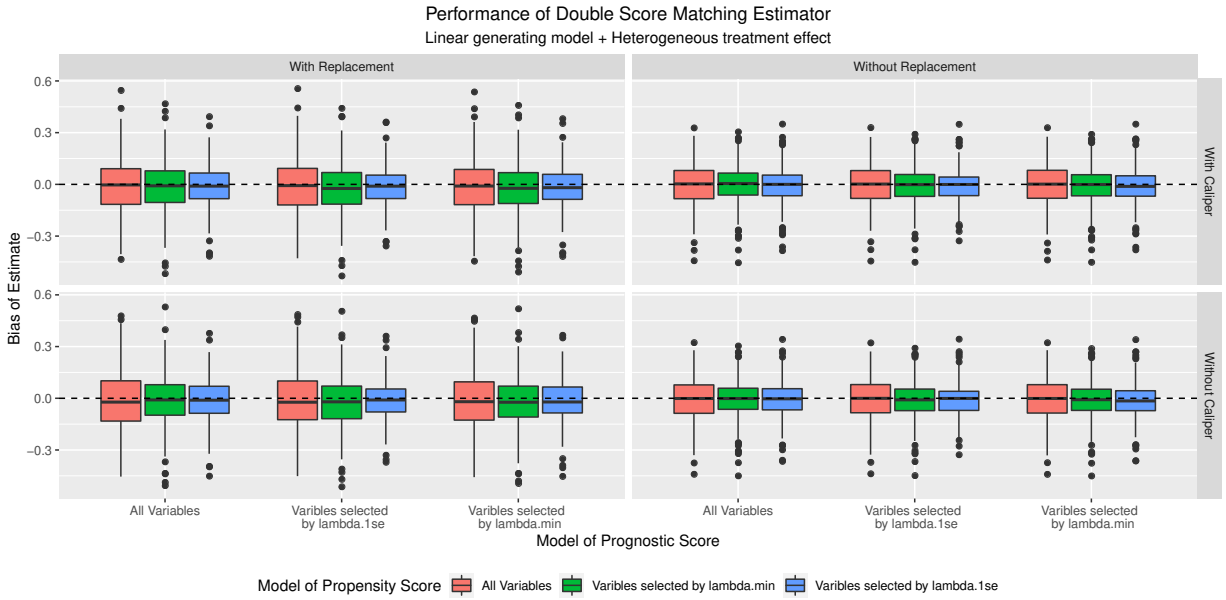


FIGURE S6 Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was linear and the treatment effect was heterogeneous.

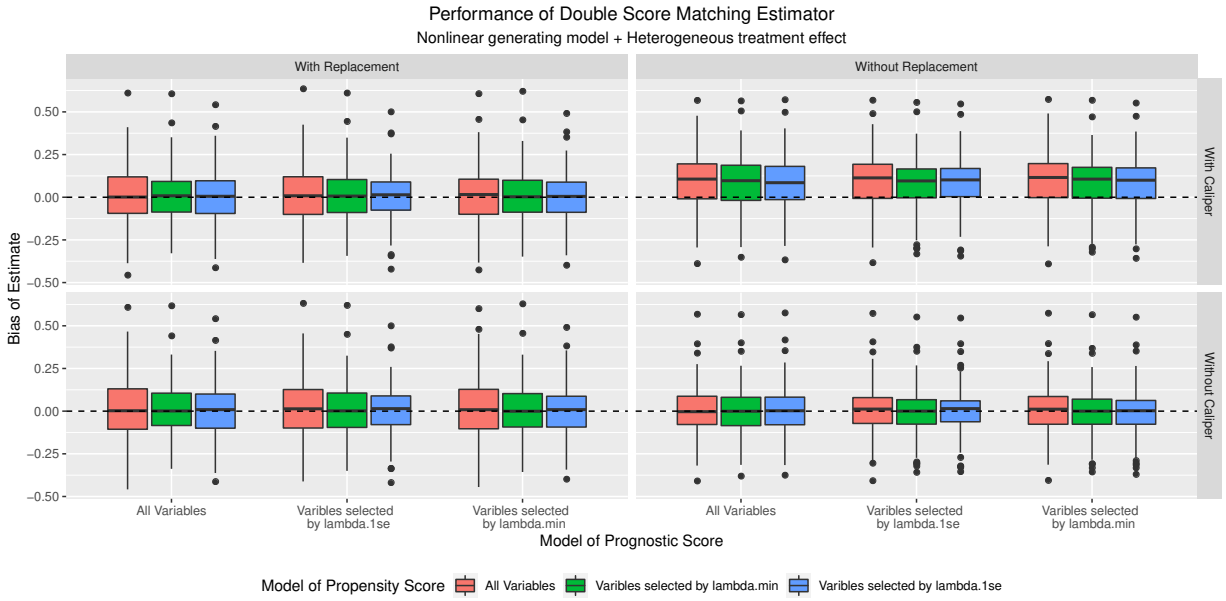


FIGURE S7 Performance of DSM estimator under different variable selection strategies in REFLECTIONS dataset. The generating model was nonlinear and the treatment effect was heterogeneous.

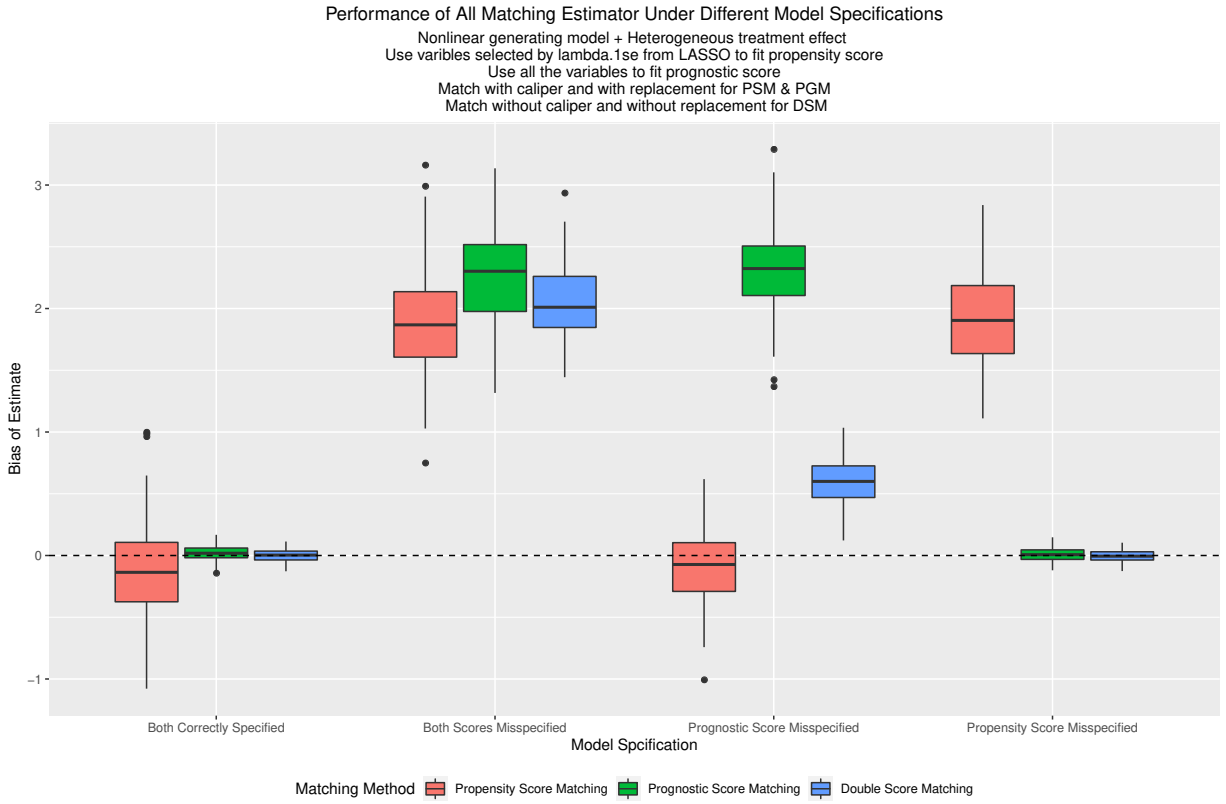


FIGURE S8 Comparison of PSM, PGM, and DSM when both models might be misspecified. DSM was applied without a caliper and without replacement.