stats

MDPI

Article

Robust Causal Estimation from Observational Studies Using Penalized Spline of Propensity Score for Treatment Comparison

Tingting Zhou [1,*], Michael R. Elliott [2] and Roderick J. A. Little [2]

[1] U.S. Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA
[2] Department of Biostatistics, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA; mrelliot@umich.edu (M.R.E.); rlittle@umich.edu (R.J.A.L.)
* Correspondence: tingting.zhou@fda.hhs.gov

Abstract: Without randomization of treatments, valid inference of treatment effects from observational studies requires controlling for all confounders because the treated subjects generally differ systematically from the control subjects. Confounding control is commonly achieved using the propensity score, defined as the conditional probability of assignment to a treatment given the observed covariates. The propensity score collapses all the observed covariates into a single measure and serves as a balancing score such that the treated and control subjects with similar propensity scores can be directly compared. Common propensity score-based methods include regression adjustment and inverse probability of treatment weighting using the propensity score. We recently proposed a robust multiple imputation-based method, penalized spline of propensity for treatment comparisons (PENCOMP), that includes a penalized spline of the assignment propensity as a predictor. Under the Rubin causal model assumptions that there is no interference across units, that each unit has a non-zero probability of being assigned to either treatment group, and there are no unmeasured confounders, PENCOMP has a double robustness property for estimating treatment effects. In this study, we examine the impact of using variable selection techniques that restrict predictors in the propensity score model to true confounders of the treatment-outcome relationship on PENCOMP. We also propose a variant of PENCOMP and compare alternative approaches to standard error estimation for PENCOMP. Compared to the weighted estimators, PENCOMP is less affected by inclusion of non-confounding variables in the propensity score model. We illustrate the use of PENCOMP and competing methods in estimating the impact of antiretroviral treatments on CD4 counts in HIV+ patients.

Keywords: causal inference; double robustness; PENCOMP; variable selection; penalized spline

Citation: Zhou, T.; Elliott, M.R.; Little, R.J.A. Robust Causal Estimation from Observational Studies Using Penalized Spline of Propensity Score for Treatment Comparison. *Stats* **2021**, *4*, 529–549. https://doi.org/10.3390/stats4020032

Academic Editor: Marco Riani

Received: 29 April 2021
Accepted: 6 June 2021
Published: 10 June 2021

1. Introduction

Observational studies are important for evaluating causal effects, especially when randomization of treatments is unethical or expensive. Valid inferences about causal effects from observational studies can only be drawn by controlling for all confounders, that is, pre-treatment variables that are related to both treatment allocation and the outcome, because the treated subjects generally differ systematically from the control subjects. For example, sicker HIV patients are more likely to take antiretroviral treatments to control the virus when their CD4 cell counts drop too low. The CD4 cell count, a measure of how well the immune system functions, is one clinical measure of the effectiveness of an antiretroviral treatment. Direct comparison of the CD4 counts between the treated and the control would lead to the false conclusion that antiretroviral treatments result in lower CD4 counts. Thus, to assess the effects of using antiretroviral treatments on the CD4 counts from an observational study, such as the Multicenter AIDS Cohort study (MACS) [1], appropriate statistical methods are needed to remove confounding by patient characteristics.

To deal with confounding by patient characteristics, the propensity score, the conditional probability of assignment to a treatment given the observed covariates, is commonly

used. Rosenbaum and Rubin (1983) [2] showed that controlling for the propensity score is sufficient to remove bias due to differences in the observed covariates between treatment groups. The propensity score summarizes the observed covariates into a single measure and serves as a dimension reduction technique. Due to the balancing property of the propensity score, the treated and control subjects with similar propensity scores can be directly compared [2]. For example, in our application, because sicker patients were more likely to be treated, we can adjust for that by controlling for the patient's probability of receiving treatment given all observed histories prior to treatment. After controlling for the propensity score, the distribution of the observed covariates, in this case, the proportion of sicker patients, will be similar between the treated and the control subjects, so the CD4 counts between the treated and the control subjects with similar propensity scores can be compared.

More generally, propensity-score-based methods first estimate the probability of treatment assignment given potential confounding variables, and then use the estimated treatment probability in weighting, or as a predictor in regression models for the outcome under alternative treatment assignments. Inverse-probability of treatment weighting (IPTW) controls for confounding by weighting subjects by the inverse of the estimated probability of receiving the observed treatment. The weights in effect create a pseudo-population that is free of treatment confounders. The IPTW estimator is consistent if the propensity score model is correct. Like IPTW, augmented IPTW estimation (AIPTW) uses the estimated propensity score as a weight but incorporates predictions from a regression model for the outcome. The AIPTW estimator consistently estimates causal effects if the propensity score model is correctly, or the outcome model is correctly specified. Both IPTW and AIPTW estimators are based on the Rubin (1974) [3] causal model framework. As such, the estimators are consistent under the causal model assumptions that there is no interference across units (stable unit treatment value assumption, SUTVA), that each unit has a non-zero probability of being assigned to either treatment group (positivity) and there are no unmeasured confounders (ignorability) [3]. Here, we mean robustness to mis-specification of the covariates in regression models, rather than robustness to outliers in the residuals. That might be achieved by replacing the assumption of normality in the distribution of errors by a longer-tailed distribution, such as Student's t [4].

Another recently developed method, Penalized Spline of Propensity Methods for Treatment Comparison (PENCOMP), imputes missing potential outcomes using regression models that include splines on the logit of the estimated probability to be assigned that treatment, as well as other covariates that are predictive of the outcome. The idea is based on the potential outcome framework of the Rubin causal model [3]. In the Rubin causal model, potential outcomes are defined as potentially observable outcomes under different treatments or exposure groups. Individual causal effects are defined as comparisons of the potential outcomes for that subject. Only the potential outcome corresponding to the treatment assigned is observed for any subject. Thus, causal inferences are based on comparisons of the imputed and the observed outcomes. Under the Rubin causal model assumptions, PENCOMP has a double robustness property for estimating treatment effects [5]. Specifically, under these standard causal inference assumptions, PENCOMP consistently estimates the causal effects if the propensity score model is correctly specified and the relationship between the outcome and the logit of the propensity score is modeled correctly, or if the relationship between the outcome and other covariates is modeled correctly.

In this paper, we study important, unresolved questions concerning how to generate robust causal inferences from observational studies. As mentioned above, common approaches to robust causal inference involve fitting two models: (a) a propensity score model, where the outcome is the indicator for which treatment is assigned and the predictors are potential confounding variables; (b) the outcome model, which relates the outcome to the treatment, and includes the propensity score as a predictor variable or as a weight,

usually the inverse of the estimated propensity to be selected. Our paper concerns practical strategies for how these regression models are specified.

For valid inferences, all true confounders should be included in the propensity score model. Ideally, we would know the set of true confounders, but in observational studies this information is rarely if ever known. Given this fact, the question of how to select variables to be included in the propensity score model is important and controversial. Some researchers have argued that all pre-treatment potential confounders should be included in the propensity model prior to seeing the outcome data, to avoid "data snooping" and mimic, as closely as possible, a randomized trial, where randomization occurs prior to observing the outcomes [6].

On the other hand, this strategy may lead to inclusion of variables that are associated with treatment selection but are not associated with the outcome and, hence, are not true confounders; including them in the propensity score model can lead to highly inefficient and non-robust inferences. The reason is the including these variables shrinks the overlap region of the propensity score distributions for the treatments, leading to weighted estimators that have extreme weights, or regression estimators that are vulnerable to model mis-specification—for example, mis-specifying a nonlinear relationship as linear. Limiting this problem argues that variable selection should consider the relationship between the variable and the outcome, provided it is not done in a way that prejudices the estimated treatment effect [7–10].

Another consideration is that including variables in the outcome model that are not associated with treatment allocation—and, hence, are not true confounders—but are related to the outcome can improve the efficiency of the causal estimate [11].

Our paper examines these aspects in detail with both simulation studies and an application, and, offers a broad discussion with a lot of important takeaways for both researchers who believe all pre-treatment confounders should be included and those who believe variable selection is always necessary. Specifically, we examine the performance of alternative confounder selection methods in PENCOMP, IPTW, and AIPTW, with and without considering the relationships between the covariates and the outcome. We also address issues of model selection and model uncertainty. For PENCOMP, we propose a new variant based on bootstrap smoothing, also called bagging. For AIPTW and IPTW, we consider an alternative approach for estimating standard errors and confidence intervals that accounts for model uncertainty.

In Section 2, we describe estimands and causal inference assumptions. In Section 3, we describe two versions of PENCOMP for estimating causal effects: one based on multiple imputation, and the other based on bootstrap smoothing, and two estimation procedures for AIPTW and IPTW. In Section 4, we describe model selection for the propensity score model and the outcome model. In Section 5, we examine using simulation studies how specification of propensity score model affects the performance of PENCOMP, AIPTW, and IPTW. In Section 6, we illustrate our methods using the Multicenter AIDS Cohort study (MACS) to estimate the effect of antiretroviral treatment on CD4 counts in HIV-infected patients. We conclude with a discussion of the results and some possible future work.

2. Materials and Methods

2.1. Estimands and Assumptions

Let X_i denote the vector of baseline covariates, and $Z_i \in \{0,1\}$ denote a binary treatment with $Z_i = 1$ for treatment and $Z_i = 0$ for control, for subject $i = 1, \cdots, N$, respectively. Under Rubin's potential outcome framework [3], causal effects at subject level are defined as the difference between the potential outcome for a subject under treatment and the potential outcome under control. Only one of the potential outcomes is observed for each subject. Let $Y_i^{Z_i}$ be the potential outcome under treatment Z_i. Here, we focus on inference about the average treatment effect (ATE), $E(Y^1 - Y^0)$, obtained by averaging subject-level causal effects across the entire population of interest.

We make the following assumptions in order to estimate causal effects.

(1) The stable unit-treatment value assumption (SUTVA) states that (a) the potential outcome under a subject's observed treatment is precisely the subject's observed outcome. In other words, there are no different versions of potential outcomes under a given treatment for each subject, and (b) the potential outcomes for a subject are not influenced by the treatment assignments of other subjects [12,13].
(2) Positivity states that each subject has a positive probability of being assigned to either treatment of interest: $0 < \Pr(Z_i = z_i \mid X_i) < 1$, where $Pr(Z_i = z_i \mid X_i)$ denotes the probability of being assigned to the treatment z_i, given the observed covariates x_i.
(3) The ignorable treatment assumption states that $(Y_i^1, Y_i^0) \perp\!\!\!\perp Z_i \mid X_i$; that is, treatment assignment is as if randomized conditional on the set of covariates X_i.

2.2. PENCOMP and Multiple Imputation

Because each subject only receives one treatment, we observe the potential outcome under the observed treatment but not the potential outcome under the alternative treatment. PENCOMP imputes the missing potential outcomes using regression models that include splines on the logit of the estimated probability to be assigned that treatment, as well as other covariates that are predictive of the outcome. We then draw inferences based on comparisons of the imputed and observed outcomes. PENCOMP, which builds on the Penalized Spline of Propensity Prediction method (PSPP) for missing data problems [14,15], relies on the balancing property of propensity score, in combination with the outcome model. Under the assumptions stated above, PENCOMP has a double robustness property for causal effects. Specifically, if either (1) the model for the propensity score and the relationship between the outcome and the propensity score are correctly specified through penalized spline, or (2) the outcome model is correct, the causal effect of the treatment will be consistently estimated [5].

Here, we briefly describe the estimation procedures for PENCOMP based on multiple imputation with Rubin's combining rules [16].

(a) For $d = 1, \cdots, D$, generate a bootstrap sample S^d from the original data S by sampling units with replacement, stratified based on treatment group. Then, carry out steps (b)–(d) for each sample S^d:
(b) Select and estimate the propensity score model for the distribution of Z given X, with regression parameters α. The estimated probability to be assigned treatment $Z = z$ is denoted as $\hat{P}_z(X) = \Pr(Z = z|X, \hat{\alpha}^d)$, where $\hat{\alpha}^d$ is the ML estimate of α. Define $\hat{P}_z^* = \log[\hat{P}_z(X)/(1 - \hat{P}_z(X))]$.
In practice, it is often unknown how treatments are assigned to subjects. There are several approaches that can be used to select the covariates to be included in the propensity score model. One approach is to include all the potential confounders from a large collection of pretreatment variables. Variables might also be selected based on how well they predict the treatment assignment. Lastly, variables can be selected based on how well they are predictive of the outcome, whether they are related to the treatment. For a binary treatment, a logistic regression is often used to model the treatment assignment.
(c) For each $z = 0, 1$, use the cases assigned to treatment group z to estimate a normal linear regression of Y^z on X, with mean

$$E(Y^z|X, Z = z, \theta_z, \beta_z) = s(\hat{P}_z^*|\theta_z) + g_z(X; \beta_z).$$

$s(\hat{P^*}_z|\theta_z)$ denotes a penalized spline with fixed knots [17–19], indexed by parameters θ_z, and $g_z()$ represents a parametric function of covariates predictive of the outcome, indexed by parameters β_z. The spline model can be formulated and estimated as a linear mixed model [19].
(d) Impute the missing potential outcomes Y^z for subjects in treatment group $1 - z$ in the original dataset S with draws from the predictive distribution of Y^z given X from

the regression in (c), with ML estimates $\hat{\theta}_z^d, \hat{\beta}_z^d$ substituted for the parameters θ_z, β_z, respectively. Repeat the above procedures to produce D complete datasets.

(e) Let $\hat{\Delta}^d$ and W^d denote the difference in treatment means and associated pooled variance estimate, based on the observed and imputed values of Y in each treatment group. The MI estimate of Δ is then $\bar{\Delta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\Delta}^d$, and the MI estimate of the variance of $\bar{\Delta}_D$

$$T_D = \bar{W}_D + (1 + 1/D)B_D, \tag{1}$$

where $\bar{W}_D = \sum_{d=1}^{D} W^d/D, B_D = \sum_{d=1}^{D}\left(\hat{\Delta}^d - \bar{\Delta}_D\right)^2/(D-1)$. The estimate Δ is t distributed with degree of freedom v, $(\Delta - \bar{\Delta}_D)T_D^{\frac{-1}{2}} \sim t_v$, where $v = (D-1)(1 + \bar{W}_D/((D+1) * B_D))^2$.

2.3. PENCOMP and Bagging

As an alternative to multiple imputation combining rules, we can draw inference about the ATE using the bagging estimator, a form of model averaging that accounts for model uncertainty. Let $S = (S_1, S_2, \cdots, S_N)$ denote the original dataset consisting of N subjects. A nonparametric bootstrap sample with replacement is denoted as $S^d = (S_1^d, S_2^d, \cdots, S_N^d)$. The procedures for PENCOMP are similar as described above, except in step (e). In step (e), the imputations are carried out on each bootstrap sample S^d, instead of the original data S.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample S^d. Repeat steps (b)–(d) for each bootstrap sample S^d to produce D complete datasets.

(b) Select and estimate the propensity score model as described in Section 2.2 (b).

(c) Estimate the outcome model Y^z on X and a penalized spline on the logit of the propensity to the treatment z using the cases assigned to treatment z.

(d) Impute the missing potential outcomes Y^z for subjects in treatment group $1 - z$ in the bootstrap sample S^d with draws from the predictive distribution estimated in (c).

(e) Let $\tilde{\Delta}$ and $\tilde{sd}_D$ denote the estimate and standard error of the causal effect, respectively. The causal estimate $\tilde{\Delta} = \sum_{d=1}^{D}\hat{\Delta}_s^d/D$, where $\hat{\Delta}_s^d$ is the mean difference in the potential outcomes obtained from bootstrap sample S^d. The standard error $\tilde{sd}_D$ is calculated as follows.

$$\tilde{sd}_D = \left(\sum_{j=1}^{N} \hat{cov}_j^2\right)^{1/2} \tag{2}$$

$$\hat{cov}_j = \sum_{d=1}^{D}(Q_{dj}^* - Q_{\cdot j}^*)(\hat{\Delta}_s^d - \tilde{\Delta})/D,$$

where $Q_{\cdot j}^* = \sum_{d=1}^{D} Q_{dj}^*/D$ and $Q_{dj}^* = \#\{S^d = S_j\}$ is the number of times that observation j of the original data S was selected into the dth bootstrap sample S^d [20]. $\hat{cov}_j$ estimates the bootstrap covariance between Q_{dj}^* and $\hat{\Delta}_s^d$. To estimate the standard error of the smoothed bootstrap causal estimate, a brute force approach would be to use a second level of bootstrapping that requires an enormous number of computations. The formula provides an approximation to such an estimate of the standard error.

Inference is made using the bootstrap smoothed estimator $\tilde{\Delta}$ and confidence interval $\tilde{\Delta} \pm 1.96\tilde{sd}_D$, instead of the Rubin's multiple imputation combining rules.

2.4. Inverse Probability Treatment Weighted Estimator IPTW

Unlike PENCOMP, IPTW does not impute potential outcomes but uses only the observed outcomes. IPTW controls for confounding by weighting subjects based on their probabilities of receiving their observed treatments. Let $\hat{P}_1(X_i, \hat{\alpha})$ denote the estimated probability of being assigned to treatment $Z_i = 1$ given the set of observed covariates $X_i = x_i$, obtained from the propensity score model for the distribution of Z_i given $X_i = x_i$, with regression parameters $\hat{\alpha}$. The treated subjects are assigned weights $1/\hat{P}_1(X_i, \hat{\alpha})$, and the control subjects are assigned weights $1/\{1 - \hat{P}_1(X_i, \hat{\alpha})\}$. Thus, the subjects who are

under-represented in a given treatment arm are given higher weights. The weights in effect create a pseudo-population where treatment groups are balanced with respective to covariate distributions. The IPTW estimator is consistent if the propensity score model is correct under the assumptions stated in Section 2.1.

The IPTW estimator is defined as

$$\hat{\Delta}_{IPTW} = \sum_{i=1}^{N} \frac{Z_i Y_i}{\hat{P}_1(X_i, \hat{\alpha})} - \sum_{i=1}^{N} \frac{(1 - Z_i) Y_i}{1 - \hat{P}_1(X_i, \hat{\alpha})}.$$

Let $\hat{\Delta}_{IPTW}$ denote the causal estimate on the original data S. Here, we consider bootstrap methods for computing its standard errors and confidence intervals. The procedures are as follows.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample S^d. Then, repeat steps (b)–(d) for each sample S^d:
(b) Select and estimate the propensity score model as described in Section 2.2 (b).
(c) Compute $\hat{\Delta}^d_{IPTW}$ for each bootstrap sample S^d.
(d) Estimate the standard error $\hat{sd}_{IPTW,D}$ for $\hat{\Delta}_{IPTW}$ based on D bootstrap samples as

$$\hat{sd}_{IPTW,D} = \sum_{d=1}^{D} (\hat{\Delta}^d_{IPTW} - \tilde{\Delta}_{IPTW})^2 / (D - 1), \tag{3}$$

where $\tilde{\Delta}_{IPTW} = \sum_{d=1}^{D} \hat{\Delta}^d_{IPTW} / D$. The standard 95% confidence intervals $\hat{\Delta}_{IPTW} \pm 1.96 \hat{sd}_{IPTW,D}$. Alternatively, the bagging estimate of the causal effect is $\tilde{\Delta}_{IPTW}$ and the 95% smoothed confidence interval is $\tilde{\Delta}_{IPTW} \pm 1.96 \tilde{sd}_{IPTW,D}$, where the smoothed standard error $\tilde{sd}_{IPTW,D}$ is computed based on Equation (2) [20].

2.5. Augmented Inverse Probability Treatment Weighted Estimator (AIPTW)

An alternative to IPTW is augmented IPTW estimation (AIPTW). AIPTW uses the estimated propensity score as a weight like IPTW but also incorporates predictions from a regression model for the outcome. Incorporating covariates predictive of the outcome in the outcome model can improve efficacy and reduce variability, especially when the weights are variable. The AIPTW estimator consistently estimates causal effects if the propensity score model or the outcome model is correctly specified under the assumptions stated in Section 2.1.

Each subject i is weighted by the balancing weight $W_i = 1 / \Big\{ Z_i P_1(X_i, \hat{\alpha}) + (1 - Z_i)(1 - P_1(X_i, \hat{\alpha})) \Big\}$. The AIPTW estimate is calculated on the original dataset S [21]:

$$\hat{\Delta}_{AIPTW} = \frac{\sum_{i=1}^{n} \{ m_1(X_i, \beta_1) - m_0(X_i, \beta_0) \}}{n} + \frac{\sum_{i=1}^{n} W_i Z_i \{ Y_i - m_1(X_i, \beta_1) \}}{\sum_{i=1}^{n} W_i Z_i} - \frac{\sum_{i=1}^{n} W_i (1 - Z_i) \{ Y_i - m_0(X_i, \beta_0) \}}{\sum_{i=1}^{n} W_i (1 - Z_i)},$$

where $m_1(X_i, \beta_1) = E(Y_i | X_i, Z_i = 1, \beta_1)$ and $m_0(X_i, \beta_0) = E(Y_i | X_i, Z_i = 0, \beta_0)$. Similar procedures as in IPTW can be used to obtain point estimates and standard 95% confidence intervals. Alternatively, the bagging estimate of the causal effect is $\tilde{\Delta}_{AIPTW}$ and the 95% smoothed confidence interval $\tilde{\Delta}_{AIPTW} \pm 1.96 \tilde{sd}_{AIPTW,D}$, can be obtained using the smoothed standard error $\tilde{sd}_{AIPTW,D}$ from Equation (2) [20].

3. Model Selection

We consider scenarios where there are some pre-treatment variables that are predictors of the outcome, some that are predictors of the treatment, some that are predictors of both the treatment and the outcome, and some that are spurious, in the sense that they

affect neither the treatment or the outcome. We consider two strategies for building the propensity score model: (1) without seeing the outcome [6], and (2) taking into account the relationships between the covariates and the outcome.

For strategy 1, one simple approach is to use the stepwise variable selection algorithm with the Bayesian Information Criterion (BIC) to select the variables that are predictive of the treatment, regardless of how well they predict the outcome. Separately, we use the same stepwise algorithm to select the outcome model for PENCOMP and AIPTW. The algorithm, abbreviated as SW, does not use outcome data and, hence, satisfies Rubin's recommendation of separating analysis from design.

In strategy 2, we use the outcome adaptive lasso approach proposed by Shortreed and Ertefaie (2017) [9]. By penalizing each covariate according to the strength of the relationship between the covariate and the outcome, the outcome adaptive lasso tends to select covariates that are predictive of the outcome and excludes covariates that are associated only with the treatment. The outcome adaptive lasso estimates for the propensity score model are:

$$\hat{\alpha}_{OAL} = argmin_{\alpha} \sum_{i=1}^{n} -Z_i(X_i^T\alpha) + log(1 + e^{X_i^T\alpha}) + \lambda_n \sum_{j=1}^{p} \hat{w}_{\alpha_j}|\alpha_j|, \quad (4)$$

where $\hat{w}_{\alpha_j} = 1/|\hat{\beta}_j|^{\gamma}$ such that $\gamma > 1$ and minimizes the mean weighted standardized difference between the treated and control. $\hat{\beta}_j$ is the coefficient estimate for covariate X_j from ordinary least square or ridge regression by regressing the outcome Y on the covariates and the treatment. Similarly, the outcome model can be selected via adaptive lasso. The adaptive lasso estimates are given as follows [22].

$$\hat{\beta}_{AL} = argmin_{\beta}||y - \sum_{j=1}^{p} X_j\beta_j||^2 + \lambda \sum_{j=1}^{p} \hat{w}_j|\beta_j|,$$

where $\hat{w}_j = 1/|\hat{\beta}_j|^{\gamma}$ and $\gamma > 0$.

This method is subject to Rubin's criticism. Excluding the treatment variable in the regressions during variable selection might reduce the potential for biasing results.

4. Simulation

We simulate each dataset as described in Zigler and Dominici (2014) and Shortreed and Ertefaie (2017) [9,23]. Each simulated dataset contains N subjects and p covariates X. The treatment Z_1 is Bernoulli distributed with logit of $P(Z_1 = 1|X) = \sum_{j=1}^{p} \alpha_j X_j$. The outcome of interest Y is normally distributed with a mean of $\eta Z_1 + \sum_{j=1}^{p} \beta_j X_j$ and a variance of 1. The treatment effect η is equal to 2, without loss of generality. We set all the coefficients 0, except the first 6 covariates $X_1, \cdots, X_6$. X_1 and X_2 are true confounders. X_3 and X_4 are predictors of the outcome only. X_5 and X_6 are predictors of the treatment only. All the other $d - 6$ covariates are spurious. We vary the strength of the relationships among the covariates, the outcome and the treatment. In the first scenario, β and α are set as: $\beta = (0.6, 0.6, 0.6, 0.6, 0, 0, 0, \cdots, 0)$, and $\alpha = (1, 1, 0, 0, 1, 1, 0, \cdots, 0)$. In the second scenario, confounders X_1 and X_2 have a weaker relationship with the treatment: $\beta = (0.6, 0.6, 0.6, 0.6, 0, 0, 0, \cdots, 0)$ and $\alpha = (0.4, 0.4, 0, 0, 1, 1, 0, \cdots, 0)$. In the third scenario, confounders X_1 and X_2 have a weaker relationship with the outcome: $\beta = (0.2, 0.2, 0.6, 0.6, 0, 0, 0, \cdots, 0)$ and $\alpha = (1, 1, 0, 0, 1, 1, 0, \cdots, 0)$. We also vary the sample sizes: N = 200 and N = 1000.

We consider four different specifications of the propensity score model: (1) True includes the true propensity score model used to generate the data; (2) trueConf includes only the true confounders; (3) outcomePred includes both the confounders and the predictors of the outcome; (4) allPoten includes all 20 variables. For these four specifications, the outcome models for PENCOMP and AIPTW are correctly specified. In addition, we

consider the following variable selection techniques for the propensity score model and the outcome model.

(a) SW: stepwise variable selection algorithm with the Bayesian Information Criterion (BIC) separately for the propensity score mode and the outcome model.
(b) OAL: outcome adaptive lasso [9] for the propensity score model, and adaptive lasso for the outcome model [22].
(c) Step-ALT: outcome adaptive lasso for the propensity score model at the first stage and then adaptive lasso for the outcome model at the second stage using only the variables that are selected at the first stage.
(d) Step-ALY: adaptive lasso for the outcome model at the first stage and then logistic regression model with all the variables selected at the first stage for the propensity score model.

We evaluate the performance of the methods based on root mean squared error (RMSE), empirical non-coverage rate of the 95% confidence interval, empirical bias, and average length of 95% confidence intervals over 500 simulated datasets. For each dataset, the standard errors and confidence intervals are estimated using 1000 bootstrap samples. Within PENCOMP, we compare the multiple imputation approach and the bagging approach. Within AIPTW and IPTW, we compare the standard approach with the bagging approach.

5. Results

Table 1 shows the results on RMSEs for sample size of 200. By comparing the four propensity score models that were fixed within each bootstrap sample: true, trueConf, outcomePred, and allPotent, we can see that excluding spurious variables or variables that were associated only with the treatment reduced the RMSEs, and including variables associated only with the outcome reduced the RMSEs. Incorporating the outcome model as in PENCOMP and AIPTW attenuated the negative effect of including nonfounding variables on RMSEs. For example, using the standard approach in scenario 1, IPTW had RMSEs of 0.19, 0.22, 0.36, and 0.41 under outcomePred, trueConf, true, and allPotent, respectively. AIPTW had RMSEs of 0.16, 0.16, 0.22, and 0.28, respectively. Using the Rubin's approach, PENCOMP had RMSEs of 0.16, 0.16, 0.21, and 0.21, respectively. Similar patterns were observed in scenario 2 and 3.

Bagging reduced the RMSEs for IPTW and AIPTW, especially when irrelevant covariates were included in the propensity score model, as in true and allPotent. The standard approach and the bagging approach yielded similar RMSEs under outcomePred and trueConf but different RMSEs under true and allPotent. For example, in scenario 1 under allPotent, IPTW had an RMSE of 0.34 when the bagging approach was used, but an RMSE of 0.41 when the standard approach was used. AIPTW had an RMSE of 0.25 when the bagging approach was used, but an RMSE of 0.28 when the standard approach was used. PENCOMP had an RMSE of 0.22 when the bagging approach was used, but an RMSE of 0.21 when Rubin's combining rule was used. For PENCOMP, the bagging approach had slightly higher RMSEs than Rubin's multiple imputation combining rule when many irrelevant variables were included as in allPotent. Similar patterns were observed in scenario 2 and 3.

The results in the variable selection cases were similar to the results without variable selection. The outcome adaptive selection procedure, such as OAL, resulted in smaller RMSEs than the variable selection procedure, such as SW, that selected variables solely based on how well they predicted the treatment. Figure A1 in the Appendix A presents the results on variable selection. For example, in scenario 1 with sample size of 200, all the variable selection procedures selected the confounders X_1 and X_2 about 99% of the time. OAL, Step-ALT, and Step-ALY selected the non-confounders X_3 and X_4 about 99% of the time, while SW selected them about 40% of the time. OAL selected X_5 and X_6 about 30% of the time; Step-ALT and Step-ALY about 8% of the time; and SW about 99% of the time.

SW, OAL, Step-ALT, and Step-ALY selected spurious variables about 40%, 34%, 8%, and 8% of the times, respectively.

An outcome adaptive selection procedure can fail to select confounders that are weakly associated with the outcome, as seen in scenario 3 in Figure A1 in the Appendix A. Similarly, the stepwise variable selection algorithm can fail to select confounders that are weakly associated with the treatment, as seen in scenario 2. Excluding weak confounders increased the bias, as seen in Table A1 in the Appendix A. However, the reduction in variance by excluding irrelevant variables when using outcome adaptive selection procedure could offset the bias and the RMSEs could still be smaller, as seen in Table 1. For example, in scenario 3, the empirical bias for IPTW was 0.146 (7%) under Step-ALT and 0.033 (2%) under SW. The RMSE for IPTW was 0.24 under Step-ALT and 0.33 under SW.

If the chosen selection procedure selects many irrelevant variables, especially the ones that are strong predictors of the treatment only, the bagging approach could reduce the RMSEs for IPTW and AIPTW. For example, in scenario 3, IPTW had an RMSE of 0.33 under SW when the standard approach was used, and an RMSE of 0.28 when the bagging approach was used. AIPTW had an RMSE of 0.25 under SW when the standard approach was used, and an RMSE of 0.23 when the bagging approach was used. PENCOMP had an RMSE of 0.21 under SW when the Rubin's combining rule was used, and an RMSE of 0.22 when the bagging approach was used. In addition, performing variable selection within each bootstrap sample could increase the chance that weak confounders were selected in some bootstrap samples. Thus, in scenario 3, the bagging IPTW and AIPTW estimators had smaller empirical biases. For example, the empirical bias for IPTW under Step-ALT was 0.146 (7%) when the standard approach was used, but 0.083 (4%) when the bagging approach was used.

Table 2 shows the results on coverage probability for sample size of 200. The bagging approach tended to have coverage rates closer to the nominal coverage than the multiple imputation approach (PENCOMP) and the standard approach (AIPTW, IPTW) for small samples. The smoothed standard errors (SE) were closer the empirical SEs so the coverage rates were closer the nominal 95% coverage, and confidence interval widths were smaller. When there were many spurious variables in the propensity score model and/or when the different models could be selected across bootstrap samples, the distribution of the bootstrap estimates could become "jumpy and erratic". Consequently, the bagging approach provided tighter confidence intervals.

As the sample size increased to 1000, the gain of using bootstrap smoothing attenuated, as seen in Tables 3 and 4. Using the standard approach of calculating the confidence intervals in the case of IPTW and AIPTW, or using multiple imputation combining rules in the case of PENCOMP, performed better than using the bagging approach. In large sample sizes, each covariate had less impact on the estimates and the selected models across the bootstrap samples were similar, so there was little variability in the bootstrap estimates. In such scenario, bagging led to greater confidence interval widths and overcoverage. In summary, bagging was advantageous when the sample size was small and the data were noisy.

Overall, both PENCOMP and AIPTW had smaller RMSEs than IPTW. PENCOMP had smaller RMSEs than AIPTW, when the propensity score model included many irrelevant covariates. Even when there was no model selection, but the sample size was small, and the propensity score model included many irrelevant variables, especially variables that were strong predictors of the treatment only, the bagging approach could stabilize the IPTW and AIPTW estimators.

Table 1. 1000× RMSE with sample size of 200. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

1000 × Empirical RMSE										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	215	190	215	278	242	278	412	313	344
Bagging	allPotent	221	196	221	251	232	251	344	294	299
Standard/Rubin	true	207	189	207	222	203	222	356	291	308
Bagging	true	207	190	207	206	193	206	329	278	286
Standard/Rubin	outcomePred	159	142	159	163	143	163	187	145	171
Bagging	outcomePred	159	142	159	161	143	161	186	145	170
Standard/Rubin	trueConf	159	143	159	161	144	161	219	186	209
Bagging	trueConf	159	143	159	159	144	159	219	186	208
Standard/Rubin	SW	214	194	213	249	231	250	382	317	327
Bagging	SW	217	196	216	230	217	230	326	280	283
Standard/Rubin	OAL	177	166	183	183	172	193	217	180	202
Bagging	OAL	178	167	184	179	168	185	206	178	193
Standard/Rubin	Step-ALT	164	149	181	165	145	234	189	147	242
Bagging	Step-ALT	164	148	181	166	149	182	189	151	188
Standard/Rubin	Step-ALY	164	148	182	164	145	236	187	146	246
Bagging	Step-ALY	164	148	182	166	149	183	190	150	191

Table 2. 1000× noncoverage rate (5%) with sample size of 200. The nominal coverage is 95%. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

1000 × Noncoverage Rate										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	8	16	8	16	16	16	14	14	8
Bagging	allPotent	34	40	34	42	52	42	60	28	40
Standard/Rubin	true	16	34	16	44	52	44	74	64	60
Bagging	true	32	36	32	34	38	34	66	62	50
Standard/Rubin	outcomePred	28	36	28	44	44	44	56	40	58
Bagging	outcomePred	24	26	24	36	32	36	48	28	48
Standard/Rubin	trueConf	30	40	30	40	46	40	54	38	34
Bagging	trueConf	24	30	24	32	42	32	38	20	30
Standard/Rubin	SW	6	18	6	10	20	10	12	12	14
Bagging	SW	32	38	32	38	54	42	60	42	46
Standard/Rubin	OAL	16	26	20	20	32	24	18	16	24
Bagging	OAL	30	40	32	38	38	36	38	30	38
Standard/Rubin	Step-ALT	24	26	32	36	36	96	40	22	106
Bagging	Step-ALT	24	24	46	38	30	56	44	32	66
Standard/Rubin	Step-ALY	24	26	34	32	34	104	36	22	108
Bagging	Step-ALY	26	22	54	36	30	58	44	30	66

Table 3. 1000× RMSE with sample size of 1000. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

1000 × Empirical RMSE										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	94	81	94	130	93	130	180	112	146
Bagging	allPotent	95	81	95	122	92	122	172	111	142
Standard/Rubin	true	94	80	94	117	89	117	186	124	157
Bagging	true	94	80	94	112	87	112	178	122	152
Standard/Rubin	outcomePred	72	64	72	74	64	74	86	65	78
Bagging	outcomePred	72	64	72	74	64	74	86	65	78
Standard/Rubin	trueConf	72	64	72	74	64	74	104	84	98
Bagging	trueConf	72	64	72	73	64	73	104	84	98
Standard/Rubin	SW	94	80	94	127	92	127	191	117	156
Bagging	SW	94	81	94	118	90	118	171	109	141
Standard/Rubin	OAL	76	67	78	78	67	81	91	69	87
Bagging	OAL	76	68	78	77	67	80	91	68	84
Standard/Rubin	Step-ALT	72	64	80	74	64	94	86	65	108
Bagging	Step-ALT	72	64	80	74	64	81	86	65	90
Standard/Rubin	Step-ALY	72	64	81	74	64	94	86	65	109
Bagging	Step-ALY	72	64	81	74	64	81	86	65	90

Table 4. 1000× noncoverage rate (5%) with sample size of 1000. The nominal coverage is 95%. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

1000 × Noncoverage Rate										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	34	44	34	50	38	50	70	54	52
Bagging	allPotent	6	10	6	10	12	10	28	10	4
Standard/Rubin	true	38	34	38	58	36	58	90	48	54
Bagging	true	4	8	4	6	12	6	28	4	10
Standard/Rubin	outcomePred	48	46	48	48	42	48	62	48	62
Bagging	outcomePred	4	2	4	8	2	8	12	6	10
Standard/Rubin	trueConf	46	48	46	52	44	52	48	60	52
Bagging	trueConf	4	0	4	6	4	6	6	6	8
Standard/Rubin	SW	34	40	34	54	44	54	76	42	54
Bagging	SW	6	8	6	8	12	8	28	6	6
Standard/Rubin	OAL	24	34	26	30	22	30	28	18	26
Bagging	OAL	4	2	4	6	4	6	6	4	6
Standard/Rubin	Step-ALT	44	36	24	48	40	60	60	44	84
Bagging	Step-ALT	4	2	2	8	4	4	12	6	6
Standard/Rubin	Step-ALY	44	36	24	48	40	56	60	44	84
Bagging	Step-ALY	4	2	2	8	4	4	12	6	6

6. Application

The Multicenter AIDS Cohort study (MACS) was started in 1984 [1]. A total of 4954 gay and bisexual men were enrolled in the study and followed up semi-annually. At each visit, data from physical examination, questionnaires about medical and behavioral history, and blood test results were collected. The primary outcome of interest was the CD4 count, a continuous measure of how well the immune system functions. We used this dataset to analyze the short term effects of using antiretroviral treatment. We restricted our analyses to visit 12. Treatment was coded as 1 if the patient reported taking any of antiretroviral treatment (ART) or enrolling in clinical trials of such drugs. We estimated the short-term (6-month) effects of using any antiretroviral treatment for HIV+ subjects. We excluded subjects with missing values on any of the covariates included in the models. We log-transformed the blood counts in this analysis.

We treated each visit as a single time point treatment. Let $t = 1$ denote the time when the treatment was administered, and $t = 2$ the time 6-month later when the outcome was measured. In addition, let $t = -1, -2, -3$ denote 1, 2, and 3 visits before the current visit $t = 1$, respectively. Let $X(t = 1, -1, -2, -3)$ denote the blood count histories prior to treatment assignment. Let Z be the binary treatment indicator. Let $Y(t = 2)$ be the CD4 count 6 months after the treatment. For the outcome model, we considered blood counts-CD4, CD8, white blood cell (WBC), red blood cell (RBC), and platelets and treatment histories from the last 4 visits. For the propensity score model, we considered the same covariates as those in the outcome model, as well as demographic variables-college education, age, and race. The treatment Z was modeled using a logistic regression. We estimated the mean CD4 count difference between the treated and the control, denoted as Δ. For PENCOMP, we replaced the simulated/imputed transformed CD4 values that were < 0 with 0 (i.e., below detection level). A total of 15 equally spaced knots and B spline were used.

Figure 1 shows that the propensity score distributions were skewed, as the treated had propensity of treatment close 1 and the control close to 0. Here, we considered the variable selection methods in the simulation studies to select the relevant variables for the propensity score model. To quantify the amount of overlap, we measured the proportion of subjects in the control group whose propensity scores were between the 95th and 5th quantiles of the propensity score distribution of the treated group, denoted as $\pi_{z=0}^{0.95} = F_{z=0}(F_{z=1}^{-1}(0.95)) - F_{z=0}(F_{z=1}^{-1}(0.05))$, where F is the cumulative distribution. Similarly, $\pi_{z=1}^{0.95}$ denotes the proportion of the treated subjects whose propensity scores were between the 95th and 5th quantiles of the propensity score distribution of the control group. Including only the covariates that were selected more than 20% of times by Step_ALT among 1000 bootstrap samples improved the overlap, as shown in Figure 1.

Table A5 in the Appendix B shows the proportion that each variable was selected across 1000 bootstrap samples. Because subjects who were treated during the recent visits were more likely to get treated again, prior treatments were highly predictive of future treatment. However, prior CD4 counts were more predictive of future CD4 count because those earlier antiretrovial treatments were not as effective. Thus, when we accounted for the outcome-covariate relationship when selecting propensity score model, prior treatment variables were selected less than 10% of the times, compared to close to 100% of the time in SW, and 58% of the time in OAL.

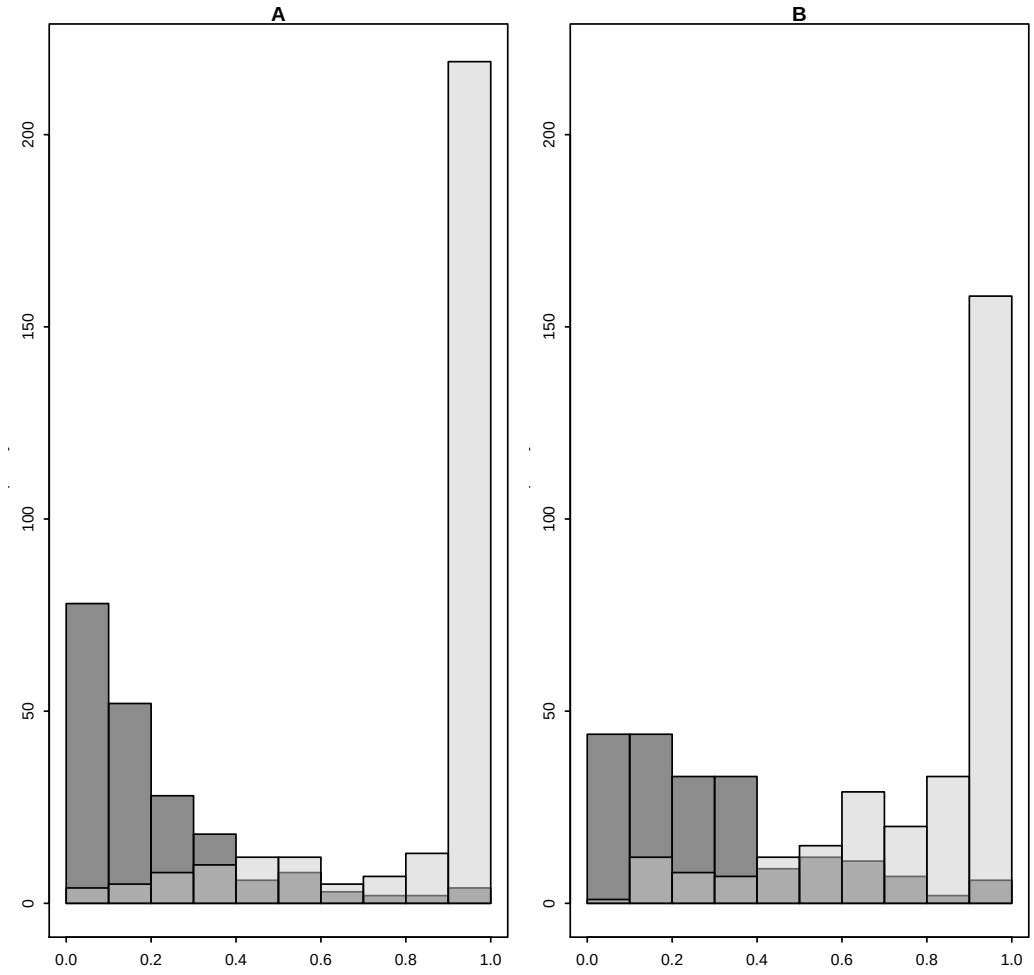

Figure 1. Propensity score distributions between the treated (grey) and control (black) if (**A**) including all covariates in the propensity score model, $\pi_{z=1}^{0.95} = 18\%$ and $\pi_{z=0}^{0.95} = 22\%$; (**B**) if including only the covariates that were selected more than 20% of times by Step_ALT among 1000 bootstrap samples, $\pi_{z=1}^{0.95} = 33\%$ and $\pi_{z=0}^{0.95} = 49\%$

We estimated the short term effect of antiretroviral treatment on CD4 count using PENCOMP, AIPTW, and IPTW, shown in Table 5. The standard errors were obtained using 1000 bootstrap samples. For PENCOMP, 1000 complete datasets were created. Overall, the IPTW estimates had the biggest confidence interval widths. Incorporating the outcome models as in AIPTW and PENCOMP decreased the standard errors and interval widths significantly. PENCOMP tended to have slightly smaller interval widths than AIPTW. The IPTW bootstrap estimates were much more variable, compared to the PENCOMP or AIPTW bootstrap estimates. As seen in the simulation studies, the bagging approach helped stabilize the IPTW and AIPTW estimators when the weights were variable. For PENCOMP, the multiple imputation approach and the bagging approach yielded similar results. Excluding irrelevant covariates from the propensity score model, as seen in Step-ALT and Step-ALY, improved the performance of IPTW significantly, in terms of the standard errors and confidence interval widths. Incorporating the outcome models in the AIPTW and PENCOMP attenuated some of the effect of including many irrelevant covariates.

Table 5. Treatment effect estimates and 95% confidence intervals.

		IPTW	AIPTW	PENCOMP
allPotent	Rubin/standard	7.5 (−2.2, 17.1)	1.3 (−0.7, 3.3)	0.7 (−1.4, 2.7)
	Bagging	5.8 (−3.0, 14.6)	0.9 (−0.9, 2.6)	0.7 (−1.0, 2.4)
SW	Rubin/standard	11.9 (1.4, 22.4)	2.7 (−0.03, 5.4)	0.9 (−1.9, 3.7)
	Bagging	6.7 (−2.7, 16.0)	1.7 (−0.5, 3.9)	0.9 (−1.3, 3.1)
OAL	Rubin/standard	2.5 (−6.6, 11.5)	0.9 (−2.1, 3.9)	0.6 (−1.5, 2.7)
	Bagging	4.9 (−3.2, 13.0)	1.6 (−0.9, 4.1)	0.6 (−1.3, 2.5)
Step-ALT	Rubin/standard	0.5 (−6.9, 7.9)	−0.4 (−2.5, 1.7)	−0.05 (−1.8, 1.7)
	Bagging	2.0 (−5.0, 9.0)	0.4 (−1.6, 2.3)	−0.04 (−1.6, 1.5)
Step-ALY	Rubin/standard	0.5 (−7.0, 7.9)	−0.4 (−2.4, 1.6)	−0.09 (−1.8, 1.7)
	Bagging	1.9 (−5.3, 9.0)	0.3 (−1.6, 2.2)	−0.08 (−1.7, 1.6)

7. Discussion

We propose a new version of PENCOMP via bagging that could improve confidence interval width and coverage, compared to PENCOMP with Rubin's multiple imputation combining rules, when the sample size is small, and the data are noisy. However, when the sample size is large and there is little variability in the bootstrap estimates, the bagging approach seems to overcover. The bagging approach and the multiple imputation approach in PENCOMP have similar RMSEs because both incorporate model selection and smooth over the estimates. Similarly, bagging improves the performance of IPTW and AIPTW in terms of RMSE, coverage and confidence interval width, especially when the sample size is small, and the data are noisy. In practice, the propensity score model is often selected, and inferences based on the selected model. This simple approach ignores model uncertainty. Compared to the standard approach for AIPTW and IPTW, the bagging approach could perform better because it incorporates model selection effects.

Our simulation studies show that excluding strong predictors of the treatment but not of the outcome, or spurious variables, helps improve the performance of the propensity score methods, especially for the weighted estimators. However, PENCOMP is less affected by inclusion of many non-confounding variables in the propensity score model than the weighted estimators because a propensity score model with many irrelevant non-confounding variables could lead to extreme propensity scores and extreme weights.

An outcome adaptive approach could help exclude strong predictors of the treatment only. However, one shortcoming of using the outcome adaptive approach is that it can miss many weak confounders. While the outcome adaptive approach can decrease the standard errors of the estimates, by excluding spurious variables and strong predictors of the treatment only, it can potentially increase bias by excluding variables that are weakly associated with the outcome, especially in small samples. This is also a shortcoming in variable selection procedures that blind the outcome, such as stepwise variable selection algorithm, because it can fail to select confounders that are weakly associated with the treatment.

Whether using an outcome adaptive approach can be beneficial depends on specific studies. When there are many weak confounders in the data, the reduction in variance from using an outcome adaptive approach might not offset the increase in bias. For example, when studying a new disease, researchers might decide to include all pretreatment variables. When there are many potential confounding variables, including those that are strongly associated with the treatment only, in the propensity score model and the weights are highly variable, PENCOMP provides a valuable approach for estimating causal effects. When variable selection is involved, smoothing over bootstrap samples can reduce the chance of excluding important confounders, which results in bias.

In high dimensional settings, including all the observed variables in the propensity score model can lead to highly unstable or even infeasible estimation. One criticism of focusing on confounders rather than just predictors of treatment assignment (i.e., balancing covariates between the treatment arms) is that incorporating the outcome in the estimation procedure, whether via prognostic score [24] or as we have done here, violates the principle that causal inference methods using observational data should mimic, as closely as possible, randomized trial designs, where outcomes are not considered until the final estimation step. Following such a rule avoids both overt and inadvertent attempts to bias model building toward preferred outcomes ("the garden of forking paths" [25], per Gelman and Loken, 2013). One approach to reducing this potential for bias is to select variables into the propensity model based a regression on the outcome that excludes variables indicating the treatments. However, with the advent of advanced "automatic" penalized regression methods, such as adaptive lasso, the risk of such "model shopping" may be sufficiently reduced, though not eliminated, so that analysts that follow the approach outlined here should endeavor to pre-specify the procedures before the analysis begins.

Author Contributions: Conceptualization, T.Z., R.J.A.L. and M.R.E.; methodology, T.Z., R.J.A.L. and M.R.E.; software, T.Z.; validation, T.Z.; formal analysis, T.Z.; investigation, T.Z.; resources, T.Z.; data curation, T.Z.; writing—original draft preparation, T.Z.; writing—review and editing, R.J.A.L. and M.R.E.; visualization, T.Z.; supervision, R.J.A.L. and M.R.E.; project administration, T.Z.; funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable as this was secondary use of publicly available data.

Informed Consent Statement: Not applicable as this was secondary use of publicly available data.

Data Availability Statement: The application datasets are publicly available. Our R code and the datasets used are available for download at https://github.com/TingtingKayla (accessed on 9 June 2021).

Acknowledgments: The authors thank the Multicenter AIDS Cohort Study (MACS) for providing us the datasets for analyses.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Additional Simulation Results

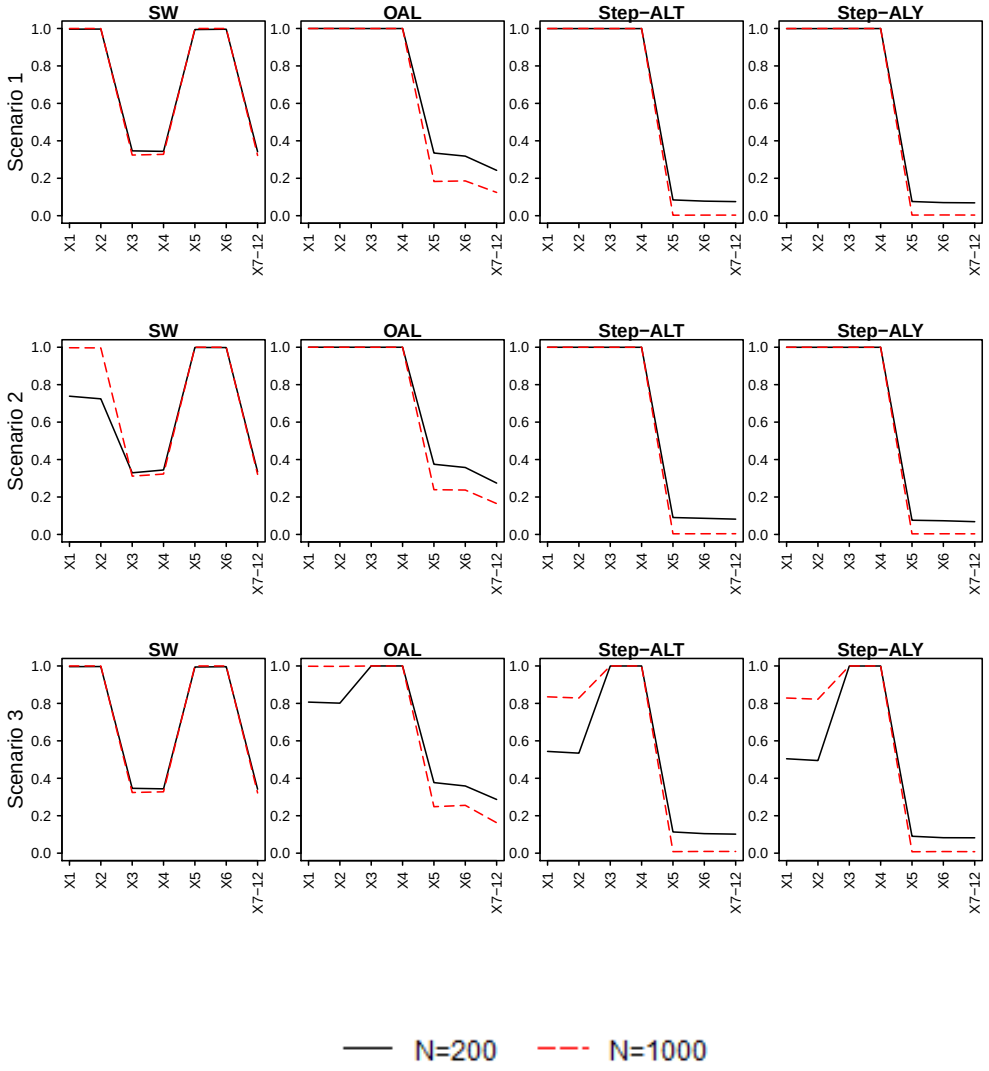

Figure A1. Proportions of each variable selected for propensity model across 500 simulated datasets and 1000 bootstrap samples for each simulated dataset for sample size of 200 and 1000. X_1 and X_2 are the true confounders; X_3 and X_4 are predictors of the outcome but not of the treatment; and X_5 and X_6 are predictors of the treatment but not of the outcome; all the other 14 covariates are spurious. Average across the spurious variables.

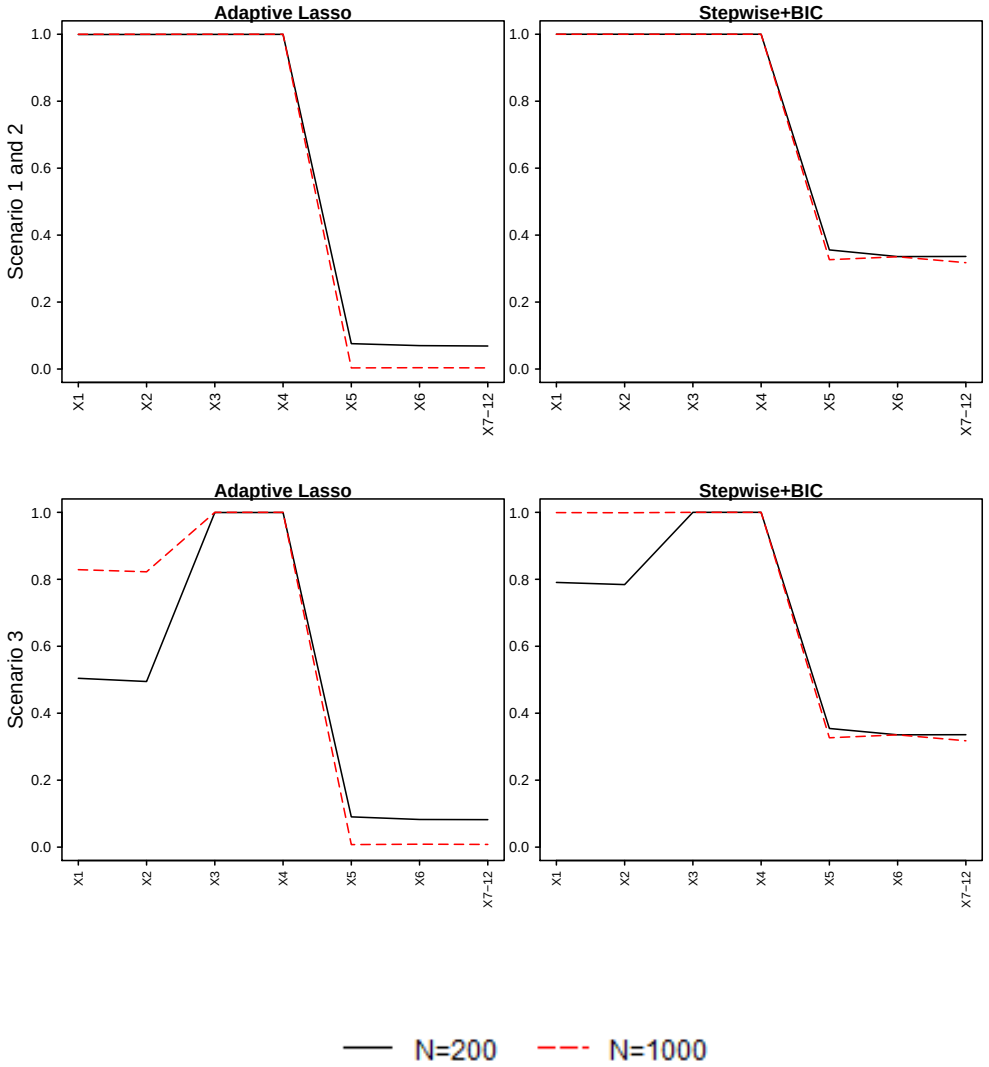

Figure A2. Proportions of each variable selected for prediction model across 500 simulated datasets and 1000 bootstrap samples for each simulated dataset for sample size of 200 and 1000. X_1 and X_2 are the true confounders; X_3 and X_4 are predictors of the outcome but not of the treatment; and X_5 and X_6 are predictors of the treatment but not of the outcome; all the other 14 covariates are spurious. Average across the spurious variables.

Table A1. 1000× empirical bias with sample size of 200. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

1000 × Empirical Bias										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	5	−2	5	2	4	2	60	11	26
Bagging	allPotent	2	−6	2	2	2	2	63	5	26
Standard/Rubin	true	6	4	6	6	3	6	61	11	27
Bagging	true	3	1	3	5	3	5	82	18	33
Standard/Rubin	outcomePred	10	7	10	8	7	8	33	9	19
Bagging	outcomePred	7	4	7	8	7	8	39	9	21
Standard/Rubin	trueConf	8	7	8	8	7	8	32	6	18
Bagging	trueConf	5	4	5	8	7	8	39	6	20
Standard/Rubin	SW	5	−2	6	4	−4	7	66	39	33
Bagging	SW	2	−5	3	1	0	5	68	37	27
Standard/Rubin	OAL	6	0	25	4	−1	17	35	2	26
Bagging	OAL	5	−1	22	2	−1	23	47	5	33
Standard/Rubin	Step-ALT	2	-4	65	7	6	132	33	8	146
Bagging	Step-ALT	3	−3	65	3	−3	66	40	2	83
Standard/Rubin	Step-ALY	2	−4	70	8	7	138	33	9	160
Bagging	Step-ALY	2	−4	70	2	−3	70	36	1	90

Table A2. 100× mean 95% confidence interval width with sample size of 200. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

100 × Mean 95% Confidence Width										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	129	101	129	131	112	131	195	158	166
Bagging	allPotent	95	81	95	98	88	98	138	117	119
Standard/Rubin	true	102	85	102	83	75	83	120	103	109
Bagging	true	92	80	92	83	76	83	118	105	110
Standard/Rubin	outcomePred	70	58	70	66	58	66	76	60	70
Bagging	outcomePred	71	63	71	69	62	69	78	63	72
Standard/Rubin	trueConf	69	58	69	65	57	65	89	76	84
Bagging	trueConf	71	63	71	69	62	69	94	82	90
Standard/Rubin	SW	129	101	128	121	104	121	180	147	154
Bagging	SW	95	81	94	92	83	92	126	107	111
Standard/Rubin	OAL	90	76	90	84	75	87	110	88	99
Bagging	OAL	78	71	79	75	70	77	89	75	82
Standard/Rubin	Step-ALT	77	66	81	73	65	78	86	69	82
Bagging	Step-ALT	73	65	75	71	64	74	80	66	75
Standard/Rubin	Step-ALY	77	65	80	73	65	78	87	68	82
Bagging	Step-ALY	72	65	75	71	64	74	80	66	75

Table A3. 1000× empirical bias with sample size of 1000. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

1000 × Empirical Bias										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	4	1	4	4	1	4	20	4	11
Bagging	allPotent	5	1	5	4	1	4	24	4	13
Standard/Rubin	true	4	2	4	5	2	5	25	5	14
Bagging	true	5	2	5	5	2	5	33	7	17
Standard/Rubin	outcomePred	0	−0	0	2	−0	2	14	1	7
Bagging	outcomePred	1	0	1	2	−0	2	16	1	7
Standard/Rubin	trueConf	0	−0	0	2	−0	2	16	0	8
Bagging	trueConf	1	0	1	2	0	2	17	1	9
Standard/Rubin	SW	4	1		3	1	3	17	4	9
Bagging	SW	5	1	5	4	1	4	25	5	13
Standard/Rubin	OAL	2	0	3	4	1	5	17	2	9
Bagging	OAL	3	1	4	3	1	5	21	3	12
Standard/Rubin	Step-ALT	1	−0	20	2	−0	22	14	1	39
Bagging	Step-ALT	0	−1	20	2	−0	21	16	1	36
Standard/Rubin	Step-ALY	1	−0	21	2	−0	23	14	1	40
Bagging	Step-ALY	0	−1	21	2	−0	22	16	1	36

Table A4. 100× mean 95% confidence interval width with sample size of 1000. The treatment effects $\eta = 2$. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

100 × Mean 95% Confidence Interval Width										
		PENCOMP			**AIPTW**			**IPTW**		
	Model Select	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**	**S1**	**S2**	**S3**
Standard/Rubin	allPotent	40	34	40	45	36	45	64	45	54
Bagging	allPotent	55	48	55	61	50	61	85	61	72
Standard/Rubin	true	40	33	40	42	35	42	63	48	55
Bagging	true	55	47	55	58	48	58	87	67	77
Standard/Rubin	outcomePred	29	25	29	29	25	29	33	26	30
Bagging	outcomePred	42	38	42	41	36	41	46	36	43
Standard/Rubin	trueConf	29	25	29	29	25	29	40	33	38
Bagging	trueConf	42	38	42	41	36	41	56	47	53
Standard/Rubin	SW	40	34	40	45	36	45	65	47	55
Bagging	SW	55	48	55	60	49	60	86	62	73
Standard/Rubin	OAL	33	30	34	34	30	35	41	32	39
Bagging	OAL	46	42	47	45	41	47	53	42	51
Standard/Rubin	Step-ALT	29	26	35	29	26	35	33	26	38
Bagging	Step-ALT	42	38	48	41	36	47	46	36	50
Standard/Rubin	Step-ALY	29	26	35	30	26	35	33	26	38
Bagging	Step-ALY	42	38	48	41	36	47	46	36	50

Appendix B. Application

Table A5. Proportion of each variable selected for prediction model across 1000 bootstrap samples.

	Outcome Model		Propensity Model			
Covariate	**SW**	**AL**	**SW**	**OAL**	**Step_ALT**	**Step_ALY**
CD4 t = −1	100	100	26	100	100	100
CD4 t = 1	100	100	100	100	100	100
CD8 t = −1	71	20	20	77	20	20
RBC t = 1	65	28	35	76	30	28
RBC t = −2	64	7	41	81	8	7
WBC t = 1	59	24	16	61	23	25
college	57	9	19	38	8	9
CD4 t = −2	52	36	19	58	32	36
platelet t = −1	49	14	37	65	12	14
CD8 t = 1	46	13	62	56	14	13
treat t = −3	43	7	38	59	6	6
treat t = −1	42	11	100	58	12	11
treat t = −2	41	7	80	42	9	7
platelet t = −3	37	4	21	38	3	4
WBC t = −1	30	1	17	40	2	1
age	24	2	28	15	1	2
CD8 t = −2	23	1	11	35	2	1
RBC t = −1	22	3	17	45	5	3
white	21	1	25	13	1	1
platelet t = 1	19	1	20	36	1	1
CD4 t = −3	18	3	12	39	3	3
CD8 t = −3	17	2	28	25	2	2
WBC t = −2	14	1	19	30	1	1
WBC t = −3	13	1	29	25	1	1
platelet t = −2	12	1	15	27	1	1
RBC t = −3	10	0	21	15	1	0

References

1. Kaslow, R.A.; Ostrow, D.G.; Detels, R.; Phair, J.P.; Polk, B.F.; Rinaldo, C.R., Jr. The Multicenter AIDS Cohort Study: Rationale, Organization, and Selected Characteristics of the Participants. *Am. J. Epidemiol.* **1987**, *126*, 310–318. [CrossRef]
2. Rosenbaum, P.R.; Rubin, D.B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
3. Rubin, D.B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J. Educ. Psychol.* **1974**, *66*, 688–701. [CrossRef]
4. Lange, K.; Little, R.J.A.; Taylor, J.M.G. Robust Statistical Modeling using the T Distribution. *J. Am. Stat. Assoc.* **1989**, *84*, 881–896. [CrossRef]
5. Zhou, T.; Elliott, M.R.; Little, R.J.A. Penalized Spline of Propensity Methods for Treatment Comparison. *J. Am. Stat. Assoc.* **2019**, *114*, 1–19. [CrossRef]
6. Rubin, D.B. The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials. *Stat. Med.* **2007**, *26*, 20–36. [CrossRef] [PubMed]
7. Brookhart, M.A.; Schneeweiss, S.; Rothman, K.J.; Glynn, R.J.; Avorn, J.; Stürmer, T. Variable Selection for Propensity Score Models. *Am. J. Epidemiol.* **2006**, *163*, 1149–1156. [CrossRef]
8. de Luna, X.; Waernbaum, I.; Richardson, T.S. Covariate Selection for the Nonparametric Estimation of an Average Treatment Effect. *Biometrika* **2011**, *98*, 861–875. [CrossRef]
9. Shortreed, S.M.; Ertefaie, A. Outcome-adaptive Lasso: Variable Selection for Causal Inference. *Biometrics* **2017**, *73*, 1111–1122. [CrossRef] [PubMed]
10. VanderWeele, T.J.; Shpitser, I. A New Criterion for Confounder Selection. *Biometrics* **2011**, *67*, 1406–1413. [CrossRef] [PubMed]
11. Rubin, D.B.; Thomas, N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics* **1996**, *52*, 249–264. [CrossRef] [PubMed]

12. Angrist, J.D.; Imbens, G.W.; Rubin, D.B. Identification of Causal Effects Using Instrumental Variables. *J. Am. Stat. Assoc.* **1996**, *91*, 444–455. [CrossRef]
13. Rubin, D.B. Discussion of "Randomization Analysis of Experimental Data: The Fisher Randomization Test" by D. Basu. *J. Am. Stat. Assoc.* **1980**, *75*, 591–593.
14. Little, R.J.A.; An, H. Robust Likelihood-Based Analysis of Multivariate Data with Missing Values. *Stat. Sin.* **2004**, *14*, 949–968.
15. Zhang, G.; Little, R.J.A. Extensions of the Penalized Spline of Propensity Prediction Method of Imputation. *Biometrics* **2009**, *65*, 911–918. [CrossRef]
16. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley: New York, NY, USA, 1987.
17. Eilers, P.H.C.; Marx, B.D. Flexible Smoothing with b-splines and Penalties. *Stat. Sci.* **1996**, *11*, 89–121. [CrossRef]
18. Ngo, L.; Wand, M.P. Smoothing with Mixed Model Software. *J. Stat. Softw.* **2004**, *9*, 1–54. [CrossRef]
19. Wand, M.P. Smoothing and mixed models. *Comput. Stat.* **2003**, *18*, 223–249. [CrossRef]
20. Efron, B. Estimation and Accuracy after Model Selection (with discussion). *J. Am. Stat. Assoc.* **2014**, *109*, 991–1007. [CrossRef]
21. Mao, H.; Li, L.; Greene, T. Propensity Score Weighting Analysis and Treatment Effect Discovery. *Stat. Methods Med. Res.* **2019**, *28*, 2439–2454. [CrossRef]
22. Zou, H. The Adaptive Lasso and its Oracle Properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
23. Zigler, c.M.; Dominici, F. Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model Averaged Causal Effects. *J. Am. Stat. Assoc.* **2014**, *109*, 95–107. [CrossRef] [PubMed]
24. Hansen, B.B. The Prognostic Analogue of the Propensity Score. *Biometrika* **2008**, *95*, 481–488. [CrossRef]
25. Gelman, A.; Loken, E. The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem Even When There Is No "Fishing Expectation" or "p-Hacking" and the Research Hypothesis Was Posited Ahead of Time. 2013. Available online: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf (accessed on 9 June 2021).