Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

______________________________ ______________________________

Xingyu Gao Date :04/26/2021

The Confounders Imbalance vs. Choices between Multiple Regression and Propensity Score Approaches

By

Xingyu Gao

Master of Public Health

Biostatistics and Bioinformatics

____

Yuan Liu, PhD

(Thesis Advisor)

____

Jeffrey Switchenko, PhD

(Reader)

The Confounders Imbalance vs. Choices between Multiple Regression and Propensity Score Approaches

By

Xingyu Gao

B.S.
University of Shanghai for Science and Technology

2019

Thesis Advisor: Yuan Liu, PhD

Reader :Jeffrey Switchenko, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in partial fulfillment of the requirements
for the degree of

Master of Public Health in Biostatistics and Bioinformatics

2021

Abstract

The Confounders Imbalance vs. Choices between Multiple Regression and Propensity Score Approaches

By Xingyu Gao

Background: The propensity score methods are widely used in observational studies as a tool for covariates balancing, especially for potential confounders. The multiple regression method and PS methods agree with each other when the baseline covariate balance is good. However, there is no clear guidance on deciding the degree of the baseline covariate balance and which method to adopt for analysis.

Methods and Materials: In this project, we created two series of simulation studies to examine the performance of PS matching with 0.2 and 0.1 calipers, ATE, ATM, and ATO PS weighting, and multiple regression under different levels of baseline covariate overlap between the two comparison groups. To create a relatively fair condition of comparison, we added two types of model misspecification to the outcome model. Instead of assessing the overlap of all covariates, we used propensity score as a summary of information. Specifically, in the simulation study, we used the overlapping coefficient (OVL) as a measurement of the degree of overlap propensity score distributions between the treatment and the control group. We evaluated the performance of different methods by absolute bias, MSE, and maximum standardized difference among all covariates related to the outcome.

Results: In the scenario that an interaction term was added in the outcome model, regardless of the strength of the interaction term or the level of model misspecification, when the OVL was above 77.0%, all methods agreed with each other. When the OVL is between 77% and 62%, ATE performed best among all methods. When OVL is below 62%, PS matching with caliper methods performed the best among all. A smaller caliper only helped to improve the matching quality when the model is almost correctly specified. ATM and ATO performed stably regardless of the OVL and the strength of model misspecification. ATO could achieve exact balance regardless of the strength of model misspecification and OVL.

Conclusion: In this paper, we propose using the OVL as a measurement of covariate balance before choosing analytic methods. When the OVL is good, multiple regression outperforms PS methods, and multiple methods can be used for cross-validation purposes. However, when the OVL is small, we proved that PS methods outperform multiple regression based on the simulation result.

The Confounders Imbalance vs. Choices between Multiple Regression and Propensity Score Approaches

By

Xingyu Gao

B.S.
University of Shanghai for Science and Technology

2019

Thesis Advisor: Yuan Liu, PhD

Reader :Jeffrey Switchenko, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in partial fulfillment of the requirements
for the degree of

Master of Public Health in Biostatistics and Bioinformatics

2021

Acknowledgement

I really appreciate the support from my family. Without their support, I could not have the chance to study at Emory and own precious friendships that I gained during the two years in Emory. I feel so honored to be a member of the Rollins School of Public Health.

I am also very grateful to Professor Yuan Liu who is my thesis advisor. Since the start of the project in 2020, she provided my guidance of research and countless and previous suggestions about academic life and career development. Working with Professor Liu is such a pleasure since her intelligence and patient guidance greatly helped the whole go through smoothly.

To the faculty and staff of the Emory Rollins School of Public Health Department of Biostatistics and Bioinformatics, thank you for your time, support, patience, and enthusiasm.

Contents

1. Introduction

1.1Background

The observational study design is commonly used to make inferences for the average treatment effect. However, in observational studies, the existence of confounding effects cannot be avoided. Therefore, achieving a covariate balance would be a necessary strategy before the outcome analysis. The propensity score is widely used as a balance score to help accomplish the covariate balancing, especially for potential confounders. The general idea of the propensity score is to reduce the high dimensions in a confounder set to a one-dimensional score, which can be viewed as a summation of all covariates (ROSENBAUM & RUBIN, 1983, p. 47). In the case of two treatment comparisons, the propensity score can be estimated by logistic regression as the estimated probability to be assigned to a treatment group. Through assessing the distribution difference in propensity score among treatment groups, we can roughly assess an overall covariate imbalance and its magnitude without examining all covariate individually. PS related methods include matching (e.g., match subjects from different treatment groups based on the similarity of their propensity score), weighting (e.g., each subject will be assigned a weight that is the inverse of the propensity score), or stratification (e.g., the treatment effect is assessed based on the stratum defined by propensity score). All of them can help create a subpopulation where the baseline covariates will be fairly balanced and hence help to eliminate the confounding effect. A propensity score method can be viewed as a strategy that identifies hidden randomized trial data from an observational database and has a strong indication of causal inference. Also, depending on the underlying covariate imbalance, the identified subpopulation might be different from the original database substantially.

As a traditional method, multiple regression is routinely used for estimating the average treatment effect or adjusted treatment effect controlling the value of potential confounders without any modification on the original study sample. However, in many studies, the multiple regression adjustment and propensity score methods give us similar results, and propensity score methods are not necessarily superior to regression adjustment (Biondi-Zoccai et al., 2011, p. 737; Elze et al., 2017, p. 352). People may be confused about the value the propensity score method can add or the gain by taking extra steps during the propensity score implementation. In this study, we try to tackle this question and address the key differences between a propensity score method and multiple regression and when they will not agree with each other.

In propensity score methods, the first step is to generate a covariate-balance population, and the second step is to evaluate treatment effect based on that pseudo population. In contrast, the multiple regression will take the original sample and control covariates in the model, which will have a similar effect as covariate-balancing under certain conditions. If the covariate-balanced sample does not differ from the original one, the two methods are highly likely to give us consistent results. This similarity between the covariate-balanced sample and the original one can be assessed by comparing the baseline covariate distribution in two populations and examining the effective sample size after the propensity score method, which is usually smaller than the original sample size after matching or weighting.

We anticipate the magnitude of imbalance or lack of overlap of underlying covariate between treatment and control groups would be the driving factor that causes the above difference. Lack of overlap will cause the estimates from PS methods to have a lower effective sample size, large variances, and low statistical power (Crump et al., 2009, p. 192). In other words, if the original population has a good covariate balance, the modification effect of PS methods is slight; thus, the estimates obtained by the PS methods will be consistent with the estimates obtained by the multiple regression. Rosenbaum and Rubin (1985b, p. 34) proposed to use the standardized difference to assess the baseline covariate balance, which is very commonly used in observational studies. Austin (2009a, p. 3098) also introduced several covariate-based balance diagnosis measurements such as graphical checking and hypothesis testing. Those methods are based on each covariate. The propensity score as an information summary can also be used to evaluate the balance of baseline covariates. In this study, we advocated using the overlapping coefficient (OVL) to quantify the degree of overlap in covariates among treatment groups. Inman and Bradley (1989, p. 3862) defined OVL as an index of the degree of agreement between two probability distributions. Belitser et al. (2011, p. 1121) introduced the method of assessing the overlap degree between the two comparison groups by calculating the OVL of the two probability density functions of the propensity score. In this way, we only need to consider this one-dimensional measurement instead of assessing all covariates. A small OVL indicates a poor overlap status and vice versa. If the OVL is very small, applying PS methods will considerably reduce the effective sample size and lead to a massive population modification. Considerable modification is also related to the issue of generalizability. In this study, we aim to explore under what overlapping circumstances that the two methods agree or disagree with each other, and how overlapping influences the performance of PS methods and multiple regression. For comparison, we used several popular matching and weighting methods in this study.

1.2 Propensity score

In 1983, Rosenbaum and Rubin (1983, p. 45) published a series of papers on propensity score analysis, in which they introduced the theory and application guidelines for a variety of propensity score models. They defined the propensity score as the conditional probability of assignment to a particular treatment given a vector of baseline covariates. Unlike an RCT, where the propensity score is predetermined, in observational studies, the propensity score of each subject is usually unknown, but we can estimate them based on the study data (Austin, 2011, p. 413). For a binary treatment, the logistic regression model can be used for estimating the propensity score for each subject. Ali et al. (2019, p. 937) advocated that for PS model selection, the propensity score estimation model should contain variables only related to the treatment but not the outcome as this may increase the variance of estimates. The estimated propensity score is the predicted probability of receiving treatment derived from the selected logistic regression model. Several methods based on propensity scores can be applied to the population for estimating the average treatment effect. This study uses the propensity score matching and weighting methods for comparison with multiple regression methods.

1.3 PS matching

The propensity score matching method is the most commonly used among all PS methods (Wang et al., 2013; Austin, 2007, p. 874980). After the propensity score of each subject has been estimated, treated subjects and untreated subjects can be matched on the propensity score.

A matched population can be built based on the estimated propensity scores. Two algorithms of matching are available: the global optimal algorithms and local optimal algorithms (Ho et al., 2007, p. 215). The Global optimal matching algorithms, introduced by Rosenbaum in 1989(1989, p. 1030), aim to minimize the total distance within the matched population. The local optimal matching algorithms aim to find the closest match for a treated subject in the control group, and it is easier to implement in practice. Based on our study design, we focused on the "one-to-one nearest neighborhood" matching method with proper calipers. The matching distance D is defined as (Rosenbaum & Rubin, 1985, p. 34):

$$D=|\text{logit }(p_1)\text{-logit }(p_2)|$$

where p_1 represents the propensity score of the treated individual and p_2 stands for the propensity score of the untreated individual in a matched pair. The matching caliper is the maximum allowed distance between each matched pair, which is defined as a specific value multiplying to the standard deviations of the logit of the propensity score (Wang et al., 2013). Lunt suggested that a tight caliper can improve the matching quality (2013, p. 227) because it considerably reduces the bias and leads to closer matches. However, Austin (2007) advocated that a caliper of 0.2 of the standard deviation of the logit of propensity score outperforms other caliper choices. Making a choice of the caliper is a trade-off matching the quality and statistical power. If the overlap is poor, a tighter caliper will lead to discarding a larger number of observations which will reduce the statistical power. In this study, we consider both loose and tight calipers for propensity score matching. Therefore, we choose 0.1 and 0.2 of the standard deviation of logit propensity scores as matching calipers.

1.4 PS weighting

The idea of weighting is to create a statistical weight based on the calculated propensity score. The distribution of propensity scores in each treatment group can be reshaped to be similar. We denote the density function of covariates X as $f(x)$. We define $w_1(x)$and $w_0(x)$as a function of x to represent the balancing weights in the treatment and control groups respectively. In this study, we adopted three types of balancing weights, ATE, ATM, and ATO. According to Li et al. (2017, p. 397), they can be expressed as follow:

ATE weight:

$$w_1(x) = \frac{1}{e(x)} \ , w_0(x) = \frac{1}{1-e(x)}$$

ATO weight:

$$w_1(x) = 1 - e(x) \ \ , w_0(x) = e(x)$$

ATM weight:

$$w_1(x) = \frac{min\{e(x), 1-e(x)\}}{e(x)} \ , w_0(x) = \frac{min\{e(x), 1-e(x)\}}{1-e(x)}$$

where e(x) represents the propensity score of each subject.

ATE weight, also known as the probability of treatment weighting (IPW), is a widely used balancing weight in practice (Li & Thomas, 2018, p. 253). The weight of the treated subject is the inverse of its propensity score, and the weight of the untreated subject is the inverse of one minus the propensity score (D’Agostino, 2007, p. 2342). As a widely used PS balancing weighting method, ATE also has the problem of sensitivity to extreme propensity scores. A large proportion of subjects with propensity scores close to 0 or 1 may lead to an extremely biased estimator with high variance (Lee et al., 2011). The superior performance of handling extreme propensity scores of ATO weight have been proven theoretically by Li et al. (2017, p. 396). According to their study, ATO weight can minimize the large sample variance and achieve exact balance if the propensity score is estimated using a logistic regression model (Li et al., 2017, p. 396). ATM weight is similar to the one-to-one pair matching without replacement method (Dehejia & Wahba, 1999, p. 1056; L. Li & Greene, 2013, p. 221).

1.5 Common support and overlapping coefficient

One of the two major assumptions of applying PS methods is the "common support" assumption defined by ROSENBAUM and RUBIN (1983, p. 54). This condition requires that all subjects have a non-zero probability of assignment to every treatment group. In practice, the presence of poor overlap in propensity score distributions between the treatment and control group in observational studies is ubiquitous. A very poor overlap status indicates the violation of this assumption.

A commonly used method of treating extreme overlap is the inverse probability trimming method (Zhou et al., 2020, p. 3726). Crump et al. (2009b, p. 194) introduced a symmetric trimming method by discarding subjects with propensity scores outside the range [0.1,0.9]. They set the cutoff points of extreme propensity scores to be 0.1 and 0.9, and the values between the two cutoff points are acceptable. This method can indeed eliminate the interference of extreme propensity scores on the estimated value, but it will also lead to loss of information. Instead of defining cutoff points of "extreme propensity score," we choose to quantify the degree of overlap status of two propensity scores distributions between the treatment group and the control group.

In this study, we assess the overlap status by using the overlapping coefficient (OVL) introduced by Inman and Bradley (1989, p. 3862). We do not know the propensity score distributions, so we use kernel density estimations for estimating the propensity score density functions (Läuter, 1988, p. 876; Belitser et al., 2011, p. 1122). The OVL is calculated using Weitzman's measure Δ, indicating the intersection area by the graphs of two propensity score probability density functions (Dhaker et al., 2017, p. 135). We define the propensity score probability density functions as f_1(x|T=1) and f_2(x|T=2). The Weitzman's measure of OVL Δ (Weitzman, 1970, pp. 1–3):

$$\Delta = \int min\{f_1(x|T = 1), f_2(x|T = 0),\}dx$$

For better understanding the modification effect of different propensity score methods under a poor overlap circumstance, we visualized propensity score distributions from the treatment group and the control group before and after applying propensity score methods in Figure 1.

For comparison, we generated two scenarios, one has a 95% OVL, and the other has a 32% OVL. Although PS weighting does not discard any data points, the sample size required to achieve the same level of precision can be calculated by effective sample size (ESS) (Bock, 2020). The ESS is calculated by:

$$\text{ESS}=\frac{N}{1+Var(w)}$$

for each treatment group separately. N represents the original population size, and w represents the weights based on the propensity score. For PS matching, the ESS represents the sample size in the matched dataset. The right column of figure 1 represents the circumstance with a poor overlap (OVL=32%), and the left column represents a circumstance with a good overlap. The mutual area of the original population for a poor overlap scenario is much smaller than that for a good overlap scenario.

A 95% OVL represents an almost "exact balance" circumstance. In this scenario, the modification effect of PS methods is relatively slight. The effective sample size after weighting and matching methods does not differ very much from the original population. But applying PS methods will also reduce the effective sample size. In this case, the population after applying PS methods will be very similar to the original population. When the OVL is 32%, the PS methods will modify the shape of the propensity score distribution, and the effective sample size becomes much less than the original sample size. In this scenario, the estimate from PS methods will differ from that of multiple regression.

Figure 1 illustrates that propensity score related approaches can help achieve similarity in the distribution of propensity score (or covariates) between treatment and control groups, but the final overall distribution would differ from the original distribution based on different PS approaches. For poor overlapped original sample (column 1), the final matched or weighted sample, except ATM or ATO weighting, could be very different from the original one (e.g., matching or ATE) with ESS substantially less than N. For a good overlapped original sample (column 2), the final matched or weighted sample by all

PS methods would not departure from the original sample too much. Also, by different PS approaches, the final sample may differ slightly among themselves.

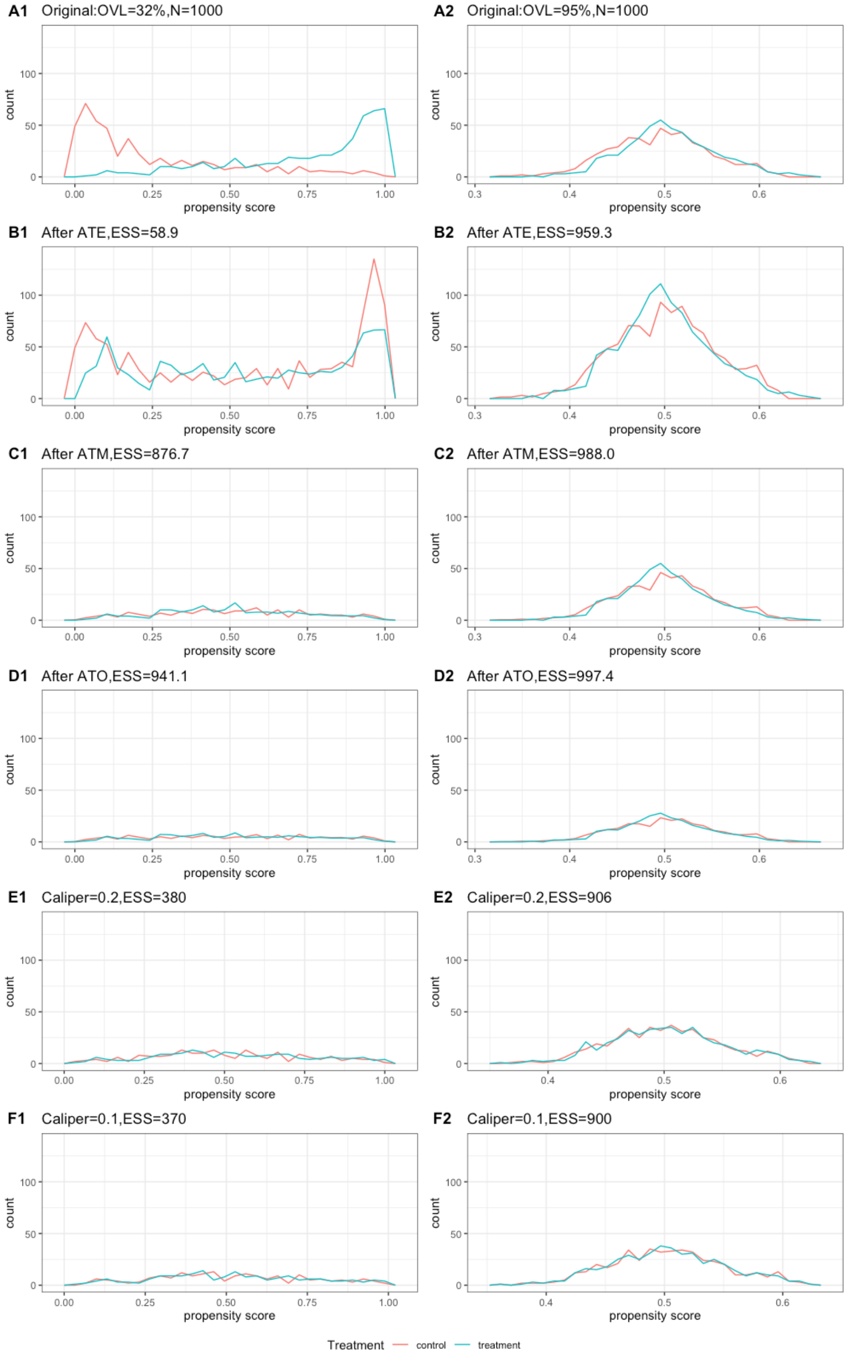

Figure 1: The distribution of propensity score in two comparison groups before and after PS methods when the OVL equals 32% and 95%. The two datasets were generated using the data generating framework in the presence of a square term in the outcome model. ESS represented the effective sample size. Picture A1 and A2 showed the propensity scores distribution before applying methods.

1.6 Standardized difference

Apart from the accuracy of treatment effect estimates, the standardized difference can be used for assessing the covariate balance after applying propensity score methods of propensity score methods. If the ignorable treatment assignment is achieved, we will observe similar propensity score distribution between each treatment group. The standardized difference can be used for covariate balance diagnosis for both continuous variables and categorical variables, and this measurement is not sensitive to the sample size (Austin, 2011, p. 413; Austin, 2009b, p. 672). For continuous covariates, the standardized difference can be expressed as follow:

$$SD = \frac{\overline{x}_{treatment} - \overline{x}_{control}}{\sqrt{\frac{s_{treatment}^2 + s_{control}^2}{2}}}$$

where $\overline{x}_{treatment}$ and $\overline{x}_{control}$ represent the sample mean of the covariate in the treatment group and the control group respectively. $s_{treatment}^2$and $s_{control}^2$ denote the sample variance in the two comparison groups.

For categorical covariates, standardized difference can be expressed as follow.

$$SD = \frac{\overline{p}_{treatment} - \overline{p}_{control}}{\sqrt{\frac{\overline{p}_{treatment}(1-\overline{p}_{treatment}) + \overline{p}_{control}(1-\overline{p}_{control})}{2}}}$$

where $\overline{p}_{treatment}$ and $\overline{p}_{control}$ represent the prevalence of the binary variable in the two comparison groups. Normand et al. (2001, p. 397) proposed that we conclude a good covariate balance if the standardized difference is smaller than 0.1. We also use the same cutoff value in our study.

2. Method

Based on our findings in Figure 1, we plan to examine how the magnitude of covariate overlapping measured as PS OVL can impact the performance among multiple regression, PS matching, and PS weighting. A simulation study was carried out.

2.1 Data generating process

We carried out two series of simulation studies to compare the relative performance of PS methods and multiple regression under different overlapping circumstances. We only considered the continuous outcome and binary treatments in this study. Since the outcome model is also generated from multiple regression models, we added model misspecifications to create a relatively fair condition for comparison. But the multiple regression model used for analysis did not contain the model misspecifications. In the first scenario, we added an interaction term between a categorical confounder and the treatment in the outcome model in the data generating process. The interaction term's existence exaggerated the difference between the target population extracted by PS methods and the original population. In this case, the treatment effect is not the same across the space of X. In the second scenario, we added a squared term of a continuous confounder to the outcome model. In this case, the true value of the treatment effect is known. After data generation, all data will be analyzed by either a PS approach or a main-effect linear regression model. For consistency, we reported the relative bias of the estimators instead of the true bias. The relative bias of the estimators is the ratio of the estimators of PS methods to that of multiple regression.

In each simulation, six covariates X_1-X_6were generated from the multivariate normal distribution with zero mean and marginal variance equals 10. We assumed that X_1-$X_4$are associated with treatment assignment, and X_3-X_6 are associated with the outcome. So X_3 and X_4 are the true confounders.

For the first scenario, we first generated six variables X_1-$X_6$from a multivariate normal distribution with zero mean and marginal variance equals ten for each data generating process. We assumed that the correlation between each pair of covariates is zero. Each covariate multiplied an index called γ to ensure that datasets with different OVL can be generated. To categorize X_3, we set two threshold values, the 1/3 quantile and the 2/3 quantile. We changed X_3 to be one if the value lies between the two thresholds; otherwise, X_3 equals zero. We then calculate the treatment assignment probability using a logistic model,

$$log(\frac{e(T)}{1-e(T)}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4$$

and simulate the treatment independently from a Bernoulli distribution.

The continuous outcome variable Y satisfied:

$$E(Y|T,X) = \beta_0 + \beta_1 X_3 + \beta_2 X_4 + \beta_3 X_5 + \beta_4 X_6 + \beta_5 X_3 * T + \Delta T$$

We chose the parameters of the treatment assignment $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$=(0,1.2,1.3,5.3,5.4). Those parameters were chosen to control the treatment prevalence to approximately 0.5. For the outcome model, we chose the parameters to be $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$=(0,2.4,2.6,2.3,1.1). The variance of Y=2 and the treatment effect $\Delta = 2$ was also fixed. The strength of the model misspecification also influenced the performance of PS methods and the multiple regression, so we considered a range of scenarios with an increasing value of β_5 to control the misspecification effect. The range of β_5 = (0,1,2,3,4,5,6,7,8). The range of γ = (0.01,0.05,0.1,0.15,0.2,0.3,0.5,1,8) so that OVL can also change approximately from 95% to 21%.

In the second scenario, we added the square term of X_3 to the outcome model. The data generating process of the six covariates the same as the previous scenario. This time, we kept the X_3 continuous.

The treatment assignment probability model was the same as the previous framework, and the treatment was also simulated from a Bernoulli distribution.

The continuous outcome variable Y satisfied:

$$E(Y|T, X) = \beta_0 + \beta_1 X_3 + \beta_2 X_4 + \beta_3 X_5 + \beta_4 X_6 + \beta_5 X_3^2 + \Delta T$$

We chose the parameters of the treatment assignment model $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$=(0,1.8,1.6,2.4,7.5). Those parameters were chosen to keep the treatment prevalence to be approximately 0.5. For the outcome model, we chose the parameters to be $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$=(0,2.4,2.6,2.3,1.1). The variance of Y = 2 and the treatment effect $\Delta = 2$ were fixed. We considered ranging β_5 = (0,1,2,3,4,5,6,7,8). In each scenario, we set γ =(0.01,0.05,0.1,0.15,0.2,0.3,0.5,1,8) so that the OVL can change approximately from 95% to 21%.

2.2 Analysis process

In each simulation, we calculated the estimators from PS weighting, matching, and multiple regression based on 1000 simulated data. For PS weighting, ATE, ATM, and ATO were used, and the treatment effect was estimated in weighted sample through a linear regression with only treatment as the only independent variable. For PS matching, we considered both caliper =0.2 and 0.1. For multiple regression, we mainly considered the model that contained X_1-X_6, but we also included the model with true confounders (X_3,X_4) and the model with all covariates (X_3-X_6) related to the outcome for reference purposes. In the summary tables, we only included the results of the model with all covariates since, in practice, we usually do not know the true confounders or the exact covariates only related to the outcome. We also calculated the maximum standardized difference of all covariates related to the outcome (X_3-X_6) before and after applying PS methods to check the covariate balance. For a specific strength of the model misspecification, we changed the value of γ to generate datasets with different OVL. For each OVL scenario, we simulated 1000 datasets with 1000 subjects in each dataset. In each generated dataset, the sample size in each comparison group is approximately 500. We calculated the mean and standard error of estimators and the maximum standardized difference for each method for the result summary. Here, we used bias and MSE to assess performance of PS methods and multiple regression. In the scenario with an interaction term in the outcome model, the true value was defined using the estimate from the multiple regression model with all covariates when the OVL is 93.3%. In the other scenario, the true value was defined as 2. The simulation study was run in R 3.6.2. The package "MatchIt" is used for PS matching methods, and "WeightIt" is used for weighting.

3. Results

Table 1: The left part of summarized average estimates and absolute bias from 1000 simulated datasets obtained using different methods in the presence of interaction terms in the outcome model with decreasing OVL. The right part summarized the MSE of estimates. The numbers of absolute bias shown in the table were represented as ten times the values. The MSE shown in the table was shown as 100 times the values. The true value was assumed as the estimate from the multiple regression model with all covariates when the OVL=93.3%. ATE, ATM and ATO represented the corresponding PS weighting schemas. Matching 0.2/0.1 represented the PS matching methods with the caliper equals 0.2/0.1.MR represents multiple regression. The category of interaction represents the value of the coefficient of the interaction term in the outcome model.

	Estimates(Absolute Bias*10)					MSE*100				
Method	93.3%	77.8%	62.0%	38.5%	21.5%	93.3%	77.8%	62.0%	38.5%	21.5%
Interaction 0										
ATE	2(0)	2(0.01)	2.01(0.04)	2.39(3.88)	4.15(21.42)	1.71	1.94	6.10	66.71	499.61
ATM	2(0)	2(0.04)	2(0.05)	1.99(0.09)	2(0.02)	1.72	1.82	2.23	3.68	7.30
ATO	2(0)	2(0.04)	2(0.06)	2(0.03)	2.01(0.06)	1.70	1.70	2.01	3.50	6.80
Matching 0.2	2.01(0.07)	2.05(0.44)	2.1(0.95)	2.17(1.67)	2.25(2.48)	2.23	2.80	4.46	9.90	19.71
Matching 0.1	2(0.02)	2.01(0.05)	2.02(0.15)	2.06(0.53)	2.13(1.23)	2.25	2.69	3.78	7.16	15.84
MR	2(0)	2(0.04)	2(0.05)	2(0.02)	2.01(0.02)	1.71	1.74	2.03	3.34	4.76
Interaction 2										
ATE	2.67(0.01)	2.67(0.01)	2.68(0.16)	3.04(3.68)	4.88(22.11)	1.53	2.07	6.12	81.30	532.48
ATM	2.66(0.05)	2.62(0.52)	2.58(0.89)	2.52(1.51)	2.47(1.95)	1.59	2.35	3.03	6.28	11.03
ATO	2.67(0)	2.64(0.29)	2.6(0.66)	2.53(1.35)	2.49(1.82)	1.54	1.98	2.52	5.41	9.86
Matching 0.2	2.67(0.01)	2.67(0.06)	2.68(0.11)	2.71(0.44)	2.75(0.86)	2.06	3.29	3.86	6.91	15.86
Matching 0.1	2.66(0.1)	2.63(0.33)	2.6(0.64)	2.58(0.92)	2.6(0.7)	2.42	3.04	4.12	8.58	14.49
MR	2.67(0)	2.64(0.31)	2.59(0.76)	2.48(1.88)	2.39(2.74)	1.54	2.02	2.66	7.03	11.92
Interaction 4										
ATE	3.34(0.02)	3.34(0.03)	3.35(0.13)	3.69(3.52)	5.67(23.31)	1.74	1.98	6.17	92.89	584.33
ATM	3.33(0.08)	3.24(0.96)	3.15(1.87)	3.03(3.09)	2.96(3.74)	1.88	2.95	6.19	13.93	22.34
ATO	3.34(0)	3.29(0.51)	3.2(1.41)	3.06(2.82)	2.98(3.56)	1.75	2.06	4.31	11.96	20.48
Matching 0.2	3.33(0.03)	3.31(0.32)	3.26(0.8)	3.25(0.84)	3.27(0.67)	2.44	3.04	5.07	8.15	17.26
Matching 0.1	3.32(0.13)	3.26(0.77)	3.18(1.58)	3.11(2.3)	3.09(2.47)	2.55	3.70	7.31	13.39	22.56
MR	3.34(0)	3.28(0.55)	3.17(1.63)	2.96(3.8)	2.8(5.42)	1.75	2.13	5.07	18.12	34.61
Interaction 6										
ATE	4(0.03)	4(0.08)	4.01(0.19)	4.39(3.97)	6.34(23.48)	1.64	1.95	6.22	94.66	602.22
ATM	3.98(0.14)	3.85(1.41)	3.73(2.59)	3.54(4.52)	3.44(5.52)	1.85	4.35	9.68	25.77	40.26
ATO	3.99(0)	3.92(0.71)	3.8(1.92)	3.58(4.1)	3.46(5.37)	1.65	2.39	6.28	21.61	37.69
Matching 0.2	3.98(0.12)	3.93(0.67)	3.86(1.31)	3.77(2.2)	3.76(2.34)	2.46	4.10	7.00	14.02	23.83
Matching 0.1	3.97(0.23)	3.88(1.16)	3.76(2.29)	3.62(3.7)	3.56(4.34)	2.80	5.14	10.35	23.29	39.17
MR	3.99(0)	3.92(0.76)	3.77(2.25)	3.43(5.65)	3.19(8.01)	1.65	2.49	7.74	36.57	69.72
Interaction 8										
ATE	4.67(0.04)	4.68(0.11)	4.69(0.26)	5.05(3.8)	7.13(24.65)	1.56	2.03	7.13	104.40	659.08
ATM	4.65(0.19)	4.48(1.85)	4.3(3.63)	4.07(6)	3.92(7.46)	1.91	6.05	16.53	41.75	67.98
ATO	4.67(0)	4.57(0.95)	4.4(2.71)	4.12(5.46)	3.94(7.24)	1.56	2.87	10.10	34.76	63.40
Matching 0.2	4.65(0.15)	4.56(1.05)	4.43(2.31)	4.33(3.35)	4.29(3.75)	2.63	5.53	11.72	22.94	37.90
Matching 0.1	4.64(0.24)	4.51(1.58)	4.33(3.34)	4.15(5.19)	4.06(6.01)	3.04	6.81	17.39	39.23	59.61
MR	4.67(0)	4.57(1.01)	4.35(3.14)	3.92(7.42)	3.58(10.86)	1.56	3.04	12.78	59.81	124.55

Note: Cells contain estimates(absolute bias*10),MSE*100

Table 1 represented the estimates, absolute bias, and MSE obtained using different methods in the presence of interaction terms in the outcome model with decreasing OVL. Adding interaction terms in the outcome model in the data generating process impacted the performance of different approaches and caused the true values of the treatment effects to be different. Therefore, we assumed the true value in each scenario to be the estimate from the multiple regression model that contained all six covariates when the

OVL is 93.3%. The "Interaction 0" scenario indicated that the multiple regression model of interest was correctly specified, and the estimate was unbiased. With the increase of the coefficient of the interaction term, all estimated values became larger. This is expected as the coefficient for interaction increases; the overall underline averaged treatment effect across all populations will increase as well. The increased strength of the interaction term and the decrease of OVL both caused more bias and MSE. Regardless of the strength of the interaction term or the level of model misspecification, when the OVL was above 77.8%, all methods seemed to agree with each other since the absolute bias for all methods was minimal (<0.05). When the interaction term existed and OVL was above 38.5%, ATE performed the best among all methods with the smallest bias and MSE regardless of the strength of the interaction term. However, ATE performed badly when the OVL was below 62%. In the "Interaction 0" scenario, we supposed the outcome model was correctly specified; hence regardless of the OVL, multiple regression always provided the most precise estimate with the smallest MSE. In this scenario, ATM and ATO performed best among all PS methods even when the overlap was very poor. Both methods had a minimal bias, and ATO had a slightly smaller MSE. For PS matching with caliper methods, a smaller caliper helped to increase the matching quality. However, when the interaction term existed, PS matching methods outperformed PS weighting methods in terms of bias. In addition, when the coefficient of the interaction term was larger than 2, a smaller matching caliper did not make the estimate more precise. The multiple regression method had a more considerable bias than ATM, ATO, and PS matching with caliper methods when the OVL was below 77.8%. When the OVL is small, PS weighting methods tend to underestimate the treatment effect while PS matching methods tend to overestimate the treatment effect.

Table 2: The left part of summarized average estimates and absolute bias from 1000 simulated datasets obtained using different methods in the presence of the square term in the outcome model with decreasing OVL. The right part summarized the MSE of those estimates. The numbers of absolute bias shown in the table were represented as ten times the values. The MSE in the table was shown as 100 times the values. The true value was 2.MR represents multiple regression. ATE, ATM, and ATO represented the corresponding PS weighting schemas. Matching 0.2/0.1 represented the PS matching methods with the caliper equals 0.2/0.1. The category of Square represents the value of the coefficient of the square term in the outcome model.

	Estimates(Absolute Bias)					MSE				
Methods	93.2%	72.2%	54.7%	33.7%	21.8%	93.2%	72.2%	54.7%	33.7%	21.8%
Square 0										
ATE	2(0)	1.99(0.06)	2.01(0.11)	2.39(3.86)	20.77(187.69)	1.71	1.95	4.13	45.37	37423.04
ATM	2(0)	1.99(0.07)	2(0.03)	2(0.01)	1.96(0.43)	1.71	1.91	2.37	4.02	38.14
ATO	2(0)	1.99(0.07)	2(0.04)	2(0.03)	2(0.04)	1.71	1.87	2.3	3.81	6.22
Matching 0.2	2(0.01)	2(0)	2.03(0.32)	2.1(0.98)	4.18(21.84)	1.78	2.23	3.25	6.76	887.33
Matching 0.1	2(0.01)	1.99(0.09)	2.01(0.09)	2.03(0.3)	2.95(9.47)	1.86	2.32	3.24	6.59	488.41
MR	2(0)	1.99(0.07)	2(0.05)	2(0.03)	2(0.03)	1.71	1.9	2.33	3.61	4.4
Square 2										
ATE	2(0.03)	2(0.01)	2.01(0.13)	2.38(3.78)	19.69(176.93)	1.5	2.06	3.84	49.82	126965.98
ATM	2(0.03)	2(0.02)	2(0.02)	1.99(0.14)	1.33(6.66)	1.49	1.99	2.43	4.77	40216.15
ATO	2(0.03)	2(0.02)	2(0.03)	1.99(0.09)	1.24(7.61)	1.5	1.95	2.34	4.58	39247.97
Matching 0.2	2(0.02)	2.01(0.11)	2.02(0.22)	2.1(1)	4.46(24.6)	1.62	2.39	3.36	7.36	56481.43
Matching 0.1	2(0.01)	2(0.04)	2(0.04)	2.02(0.22)	2.42(4.24)	1.64	2.36	3.1	7.21	57185.31
MR	2(0.03)	2(0.03)	2(0.04)	2(0.04)	1.66(3.42)	1.5	1.97	2.4	4.18	37286.79
Square 4										
ATE	2(0.04)	1.99(0.05)	2.01(0.12)	2.35(3.53)	21.29(192.92)	1.64	1.87	3.51	67.19	379826.61
ATM	2(0.04)	2(0.04)	2.01(0.06)	1.99(0.08)	4.14(21.44)	1.63	1.81	2.34	5.42	180140.53
ATO	2(0.04)	2(0.04)	2.01(0.07)	1.99(0.06)	4.21(22.06)	1.64	1.79	2.22	5.17	173354.06
Matching 0.2	2(0.05)	2(0.04)	2.03(0.32)	2.1(1.04)	4.98(29.82)	1.78	2.12	3.09	9.41	233250.48
Matching 0.1	2(0.03)	2(0.04)	2.02(0.16)	2.04(0.4)	5.06(30.58)	1.78	2.22	2.98	8.06	247073
MR	2(0.04)	2(0.04)	2.01(0.08)	2(0.01)	4.42(24.22)	1.65	1.82	2.24	5.18	144909.56
Square 6										
ATE	2(0)	2(0.01)	2.02(0.16)	2.38(3.8)	24.45(224.55)	1.55	2.01	4.26	52.62	614118.19
ATM	2(0)	2(0.03)	2(0.03)	1.98(0.17)	3.2(12.03)	1.54	1.91	2.53	7.21	363446.21
ATO	2(0)	2(0.02)	2(0.04)	1.99(0.11)	3.69(16.92)	1.55	1.9	2.46	6.78	348768.73
Matching 0.2	2(0.01)	2.01(0.06)	2.03(0.32)	2.1(1)	5.31(33.11)	1.65	2.26	3.58	11.88	468301.74
Matching 0.1	2(0.02)	2(0.01)	2.01(0.08)	2.02(0.2)	4(19.99)	1.7	2.28	3.38	10.55	474258.76
MR	2(0)	2(0.02)	2(0.03)	1.98(0.19)	4.51(25.14)	1.55	1.93	2.5	6.92	328804.71
Square 8										
ATE	2(0.01)	2(0.03)	2.01(0.11)	2.41(4.07)	18.22(162.18)	1.79	2.13	3.6	66.48	1271548.41
ATM	2(0.01)	2(0.04)	2(0)	2(0.02)	0.16(18.39)	1.79	2.02	2.51	10.6	668927.33
ATO	2(0.01)	2(0.04)	2(0.01)	2.01(0.06)	1.04(9.62)	1.79	2	2.45	10.34	644338.12
Matching 0.2	2(0.01)	2(0.03)	2.03(0.25)	2.12(1.23)	0.38(16.22)	1.94	2.38	3.4	16.15	912964.54
Matching 0.1	2(0.01)	2(0)	2(0.04)	2.04(0.44)	-0.51(25.12)	1.97	2.42	3.35	15.39	918054.88
MR	2(0.02)	2(0.04)	2(0.01)	2(0)	2.75(7.53)	1.79	2.03	2.53	10.19	596244.82

Note: Cells contain estimates(Absolute Bias*10),MSE*100

Table 2 represented the estimates, absolute bias, and MSE obtained using different methods in the presence of squared terms in the outcome model with decreasing OVL. The “Square 0” scenario indicated that the multiple regression model of interest is correctly specified, and the estimate was unbiased. Adding squared terms for one of the confounders in the outcome model in the data generating process did not impact the estimation of the true treatment effect. If we discarded the scenario of OVL equals 21.8%, the general results of the estimate were similar to the results from Table 1. All PS methods seemed to agree with multiple regression when the OVL was larger than 33.7%. However, when the OVL was below 21.8%, all methods give divergent results. It seemed like the strength of the square term did not impact MSE if we discard the scenario with 21.8% OVL. When the OVL reduced from 33.7% to 21.8%, MSE from all methods sharply increased. When OVL was below 54.7%, as the square term's intensity increases, the MSE of all methods increases drastically.

The maximum standardized difference of covariate related to the outcome was summarized in

Table 3 in the appendix to demonstrate the covariate balance of different PS methods. Regardless of the strength of the model misspecification, the maximum standardized difference became larger with the decrease of OVL. The strength of model misspecification also had very little impact on this increase. In the scenario with the interaction term in the outcome model, ATM and PS matching with caliper =0.1 all had small maximum standardized difference. ATO achieved the exact balance regardless of the value of OVL and the strength of model misspecifications in both scenarios.

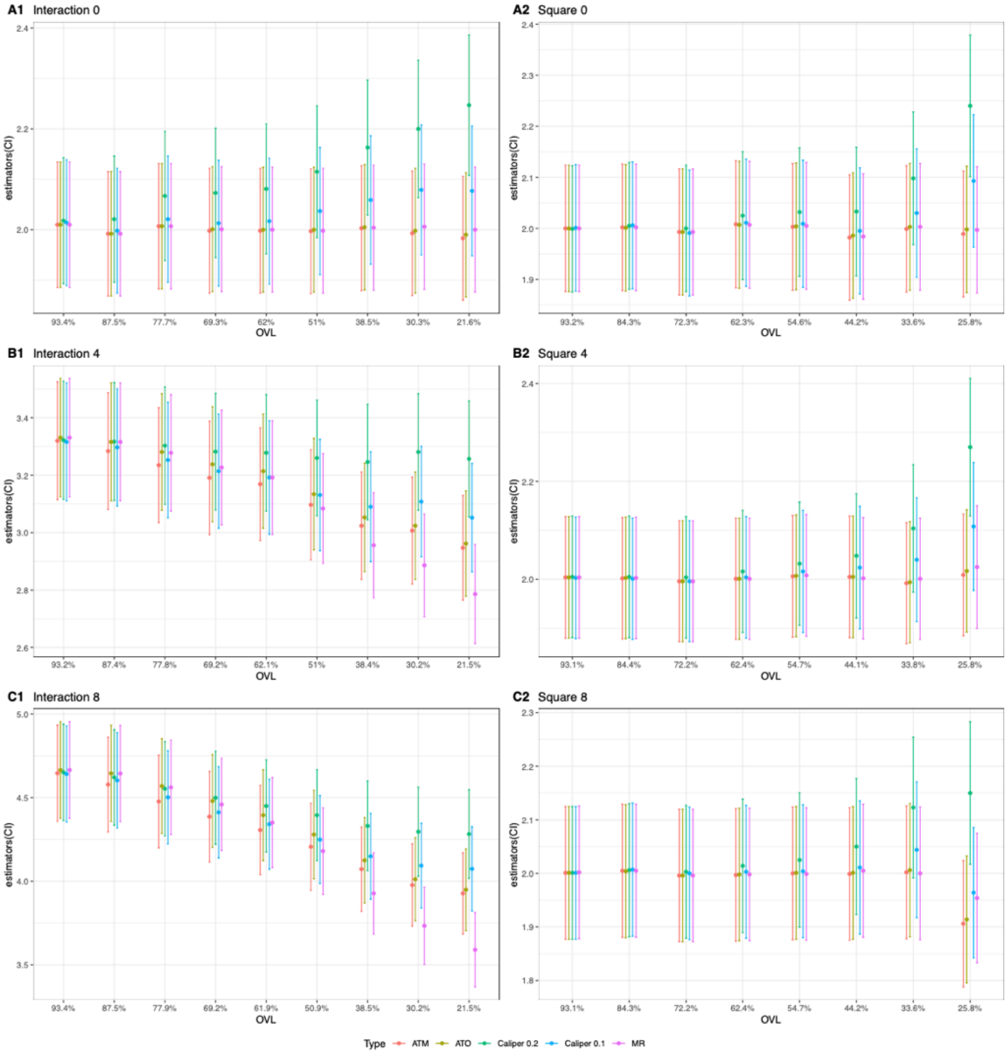

Figure 2: 95%CI of estimators with different coefficients of model misspecification.

For a better understanding of the study result, Figure 2 provides an overall visualization of trend change under different level of model misspecification by multiple regression, ATM, ATO, and matching with caliper=0.In the scenario with an interaction term in the outcome model, which was indicated the left part of the figure, the four methods gradually diverged with the decreasing of OVL. When the OVL was very small, the multiple regression estimators tended to underestimate treatment effect than PS methods estimators, while the PS matching with caliper method usually gave relatively large estimators. ATM and ATO methods were always consistent regardless of the strength of the interaction term and the overlapping status. Multiple regression estimators tended to have a significant difference between the estimators from PS methods when the OVL was smaller than 38.5%. The existence of the interaction term could further drive the bias. In the scenario with a square term in the outcome model, the four methods gave similar results when the OVL was above 33.8%, regardless of the strength of the squared term in the outcome model. In this scenario, the true value of the treatment is 2. The bias increased as OVL decreased in the first scenario. The trend was the same in the second scenario but not as extreme as the first.

4. Discussion

In this study, we created a straightforward data structure to compare the performance of several PS matching and weighting methods and multiple regression. We chose to use a dichotomous treatment variable and a normally distributed outcome. In each simulated dataset, only six covariates were included without any missing data and unmeasured confounders. The data generating process were similar to a study design procedure of an RCT study. The true propensity score model contained X_1-X_4, which were related to the treatment. The treatment assignment procedure was based on the true propensity score for each subject. The outcome satisfied a multiple regression model and was also simulated using the normal distribution. To create a relatively fair condition for PS methods and multiple regression, we added model misspecification to the outcome model, and the multiple regression model used for analysis only contained all six covariates. We considered two types of model misspecification. In the first scenario, an interaction term of a categorical confounder was added in the outcome model. This setting could mimic a heterogeneous treatment effect among different subpopulations. The true treatment effect in the original population was known, so the relative performance of different methods can be compared. In the second scenario, a square term of a continuous confounder was added. This setting imitated one of the practical concerns of an observational study-- an unknown true outcome model. In the data analysis process, we adopted the variable set's inclusion criteria mentioned in Austin's (2011, p. 412) study. The selected propensity score model only contained potential confounders (covariates related to the outcome but not the treatment, in our study design, it is X_3-X_6). In our simulation, the degree of overlap could be changed by controlling the distance between each subject within a simulated dataset. An exaggerated distance led to a dispersed covariate distribution. Furthermore, the degree of the disperse was quantified by OVL.

According to our simulation study results, we found that the influence of OVL on the degree of divergence was not continuous. However, when OVL was less than a specific value, as OVL continued to decrease, the estimates of multiple regression would diverge from the estimates of PS methods. For different PS methods, this cutoff point was different. In other words, regardless of the type of model misspecification, when the overlap was relatively good, all methods tended to agree with each other. The estimates from the PS methods and those from the multiple regression disagreed with each other only when the OVL was very small. The strengths of model misspecification did not seem to impact this trend but the magnitude of the difference between PS methods and multiple regression. The standard error of estimators from all PS methods became larger with the decrease of OVL.ATE was the most sensitive to the decline of OVL and the strength of the model misspecification. PS matching with caliper methods were also sensitive to the reduction of OVL but not as extreme as ATE. The OVL could act as a factor when considering the matching caliper. A narrow caliper outperformed a wide one when the OVL is relatively poor. ATM and ATO performed stably and tended to be more consistent with multiple regression. When applying PS methods under a poor overlap, the effective sample size used for analysis would be much smaller than the original population.

Our study design and results were very similar to the theory proposed in Austin's (2011, p. 411) study. When satisfying "no unmeasured confounders," "continuous outcome," and "correctly specified propensity score model and outcome model," the results from multiple regression and PS methods would coincide. Hade and Lu (2013, p. 81) set three covariate distribution shapes and named them as "contained," "common support," and "some overlap" in their simulation study. The three types had a decreasing degree of overlap but no extreme overlap circumstances. F. Li and Thomas (2018, p. 254) generated four different propensity overlap scenarios in their simulation study, and they also considered extreme overlap circumstances. Inspired by their studies, we advocated that quantifying the overlapping degree using OVL can better help balance diagnostics before adopting any analytic method. In practice, the true propensity score model and the outcome model are unknown in observational studies. Therefore, the major practical concern is model selection. Once we select a proper propensity score model, the OVL can be calculated to assess the degree of the baseline covariate overlap. The decision of the next step can be made by the research question and OVL. If we are confident enough to claim that the prediction model is correctly specified, the multiple regression model is the best choice. When the OVL is above 90%, multiple regression and PS methods agree with each other. Multiple methods can be adopted for cross-validation purposes. ATE performed very well when the OVL is above 38%. When the OVL is below 62%, propensity score matching with caliper methods, ATO, and ATM could all be selected based on the research question.

Our simulation study has several limitations. In this study, we only considered a binary treatment and a continuous outcome. As mentioned in Austin's (2011, p. 411) study, one of the advantages of propensity score methods is the flexibility of handling various types of outcomes such as time-to-event outcomes and a binary outcomes. It is also of interest to explore whether the results can be expanded to other scenarios. In the scenario with a square term of confounder in the outcome model, the pattern of the performance of different methods was not very clear. It is probably because the confounder's strength in the outcome model in the data generating process was relatively small.

5. Conclusion

In conclusion, we suggest estimating the propensity score of each observation and assessing the baseline covariate balance by OVL and PS methods outperform multiple regression when the overlap is poor. According to our study result, PS methods and multiple regression agree when the OVL is good. However, PS methods provided more efficient estimates when the OVL is poor. PS matching with a smaller caliper does not always improve the matching quality. This improvement can be achieved only when the model is correctly specified. ATE has superior performance when the OVL is good, but it should be avoided when the overlap is very poor. The decision to adopt the analytic method should be made after estimating the OVL according to the research question.

Reference

Ali, M. S., Prieto-Alhambra, D., Lopes, L. C., Ramos, D., Bispo, N., Ichihara, M. Y., Pescarini, J. M., Williamson, E., Fiaccone, R. L., Barreto, M. L., & Smeeth, L. (2019). Propensity Score Methods in Health Technology Assessment: Principles, Extended Applications, and Recent Advances. *Frontiers in Pharmacology*, *10*, 937. https://doi.org/10.3389/fphar.2019.00973

Austin, P. C. (2007). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, *134*(5), 1128–1135.e3. https://doi.org/10.1016/j.jtcvs.2007.07.021

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107. https://doi.org/10.1002/sim.3697

Austin, P. C. (2009b). The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. *Medical Decision Making*, *29*(6), 661–677. https://doi.org/10.1177/0272989x09341755

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H. H., de Boer, A., & Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety*, *20*(11), 1115–1129. https://doi.org/10.1002/pds.2188

Biondi-Zoccai, G., Romagnoli, E., Agostoni, P., Capodanno, D., Castagno, D., D’Ascenzo, F., Sangiorgi, G., & Modena, M. G. (2011). Are propensity scores really superior to standard multivariable analysis? *Contemporary Clinical Trials*, *32*(5), 731–740. https://doi.org/10.1016/j.cct.2011.05.006

Bock, T. (2020, December 7). *What is Effective Sample Size?* Displayr. https://www.displayr.com/what-is-effective-sample-size/

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

D’Agostino, R. B. (2007). Propensity Scores in Cardiovascular Research. *Circulation*, *115*(17), 2340–2343. https://doi.org/10.1161/circulationaha.105.594952

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858

Dhaker, H., Ngom, P., & Mbodj, M. (2017). Overlap coefficients based on Kullback-Leibler divergence: Exponential populations case. *International Journal of*

Applied Mathematical Research, *6*(4), 135. https://doi.org/10.14419/ijamr.v6i4.8493

Elze, M. C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G. W., & Pocock, S. J. (2017). Comparison of Propensity Score Methods and Covariate Adjustment: Evaluation in 4 Cardiovascular Studies. *Journal of the American College of Cardiology*, *69*(3), 345–357. https://doi.org/10.1016/j.jacc.2016.10.060

Hade, E. M., & Lu, B. (2013). Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine*, *33*(1), 74–87. https://doi.org/10.1002/sim.5884

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Inman, H. F., & Bradley, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods*, *18*(10), 3851–3874. https://doi.org/10.1080/03610928908830127

Läuter, H. (1988). Silverman, B. W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall, London – New York 1986, 175 pp., £12.—. *Biometrical Journal*, *30*(7), 876–877. https://doi.org/10.1002/bimj.4710300745

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight Trimming and Propensity Score Weighting. *PLoS ONE*, *6*(3), e18174. https://doi.org/10.1371/journal.pone.0018174

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2017). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, *113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Li, F., & Thomas, L. E. (2018). Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*, 250–257. https://doi.org/10.1093/aje/kwy201

Li, L., & Greene, T. (2013). A Weighting Analogue to Pair Matching in Propensity Score Analysis. *The International Journal of Biostatistics*, *9*(2), 215–234. https://doi.org/10.1515/ijb-2012-0030

Lunt, M. (2013). Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance With Propensity Score Matching. *American Journal of Epidemiology*, *179*(2), 226–235. https://doi.org/10.1093/aje/kwt212

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly. *Journal of Clinical Epidemiology*, *54*(4), 387–398. https://doi.org/10.1016/s0895-4356(00)00321-8

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

ROSENBAUM, P. A. U. L. R., & RUBIN, D. O. N. A. L. D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032. https://doi.org/10.1080/01621459.1989.10478868

Rosenbaum, P. R., & Rubin, D. B. (1985a). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383

Rosenbaum, P. R., & Rubin, D. B. (1985b). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383

Samawi, H. M., Helu, A., & Vogel, R. (2011). A nonparametric test of symmetry based on the overlapping coefficient. *Journal of Applied Statistics*, *38*(5), 885–898. https://doi.org/10.1080/02664761003692365

Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, J., & Xia, J. (2013). Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study. *PLoS ONE*, *8*(12), e81045. https://doi.org/10.1371/journal.pone.0081045

Weitzman, M. S. (1970). *Measures of Overlap of Income Distributions of White and Negro Families in the United States*. U.S. Bureau of the Census.

Zhou, Y., Matsouaka, R. A., & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, *29*(12), 3721–3756. https://doi.org/10.1177/0962280220940334

Appendix:

Table 3: Maximum standardized difference of all covariates related to the outcome after PS methods in presence of the interaction /square term in the outcome model with decreasing OVL.MR represents multiple regression. The category of Int/Sq represents the value of coefficient of the interaction term in the outcome model.

	Interaction					Square				
Methods	93.3%	77.8%	62.0%	38.5%	21.5%	93.2%	72.2%	54.7%	33.7%	21.8%
Int/Sq 0										
Multiple regression	0.11	0.51	0.91	1.53	1.94	0.12	0.69	1.17	1.79	2.11
ATE	0	0.01	0.06	0.28	0.95	0	0.03	0.1	0.42	1.01
ATM	0	0.01	0.01	0.02	0.03	0	0.01	0.02	0.02	0.03
ATO	0	0	0	0	0	0	0	0	0	0
Matching 0.2	0.02	0.04	0.05	0.09	0.14	0.02	0.04	0.07	0.11	0.15
Matching 0.1	0.02	0.04	0.05	0.07	0.11	0.03	0.04	0.06	0.08	0.11
Int/Sq 2										
Multiple regression	0.11	0.51	0.91	1.53	1.94	0.12	0.69	1.18	1.79	2.12
ATE	0	0.01	0.06	0.28	0.96	0	0.03	0.1	0.42	1.02
ATM	0	0.01	0.01	0.02	0.03	0	0.01	0.02	0.02	0.03
ATO	0	0	0	0	0	0	0	0	0	0
Matching 0.2	0.02	0.04	0.05	0.09	0.14	0.02	0.04	0.06	0.11	0.15
Matching 0.1	0.02	0.04	0.05	0.07	0.1	0.03	0.04	0.06	0.08	0.11
Int/Sq 4										
Multiple regression	0.11	0.51	0.92	1.52	1.95	0.12	0.69	1.17	1.79	2.12
ATE	0	0.02	0.06	0.29	0.95	0	0.02	0.1	0.44	1.02
ATM	0	0.01	0.01	0.02	0.03	0	0.01	0.02	0.02	0.03
ATO	0	0	0	0	0	0	0	0	0	0
Matching 0.2	0.02	0.04	0.05	0.09	0.15	0.02	0.04	0.06	0.1	0.15
Matching 0.1	0.02	0.04	0.05	0.07	0.1	0.03	0.04	0.06	0.08	0.11
Int/Sq 6										
Multiple regression	0.11	0.51	0.91	1.53	1.94	0.12	0.69	1.18	1.79	2.12
ATE	0	0.01	0.06	0.28	0.95	0	0.03	0.1	0.41	1.02
ATM	0	0.01	0.01	0.02	0.03	0	0.01	0.02	0.02	0.03
ATO	0	0	0	0	0	0	0	0	0	0
Matching 0.2	0.02	0.04	0.05	0.09	0.14	0.02	0.04	0.06	0.11	0.15
Matching 0.1	0.02	0.04	0.05	0.07	0.1	0.03	0.04	0.06	0.08	0.11
Int/Sq 8										
Multiple regression	0.11	0.51	0.91	1.52	1.94	0.12	0.69	1.17	1.8	2.11
ATE	0	0.01	0.06	0.29	0.95	0	0.02	0.09	0.41	1.01
ATM	0	0.01	0.01	0.02	0.03	0	0.01	0.02	0.02	0.03
ATO	0	0	0	0	0	0	0	0	0	0
Matching 0.2	0.02	0.04	0.05	0.09	0.14	0.02	0.05	0.06	0.11	0.15
Matching 0.1	0.02	0.04	0.05	0.07	0.1	0.03	0.04	0.06	0.08	0.11