Journal of Data Science, 18(2). P. 339 - 353,2020
DOI:10.6339/JDS.202004_18(2).0007

THE PERFORMANCE OF LARGEST CALIPER MATCHING THROUGH MONTE CARLO SIMULATION AND AN APPLICATION TO SUPPORT DATA

Sharif Mahmood

Department of Mathematics, University of Central Arkansas

Abstract

The paper presents an investigation of estimating treatment effect using different matching methods through Monte Carlo simulation. The study proposed a new method which is computationally efficient and convenient in implication—*largest caliper* matching and compared the performance with other five popular matching methods. The bias, empirical standard deviation and the mean square error of the estimates in the simulation are checked under different treatment prevalence and different distributions of covariates. It is shown that largest caliper matching improves estimation of the population treatment effect in a wide range of settings compare to other methods. It reduces the bias if the data contains the selection on observables and treatment imbalances. Also, findings about the relative performance of the different matching methods are provided to help practitioners determine which method should be used under certain situations. An application of these methods is implemented on the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) data and, important demographic and socioeconomic factors that may affect the clinical outcome are also reported in this paper.

Key words and phrases: Covariate balance, bias reduction, largest caliper matching, matching methods, treatment effect.

1. Introduction

Matching methods are popular to estimate the unbiased estimate of the treatment effect both in randomized and non-randomized experiments. In randomized experiment, researchers use matching methods to form pair/block similar subjects and assign treatments. In non-randomized experiment, researchers use pretreatment covariates to match the treated subjects with control subjects and attempt to replicate a randomized experiment as if the treatments were randomly assigned. When the covariate distributions of the treated and control subjects are different—crude analysis could make a substantial bias. This bias can be crucial in causal inference for any non-randomized experiment because it is systematically generated and structured by covariates which influence the cause and/or its related outcomes (Kang et al, 2012). An appropriate matching method should reduce bias due to covariates by reducing the observed and unobserved covariate imbalances between treated and control groups.

There are plenty of matching methods that have been developed in literature that improve the covariate balance iteratively by estimating a distance between treated units and potential controls, finding the matches, and checking balance until a satisfactory level is achieved. When there are large number of covariates—there is a potential risk of confounding and it may not possible to reduce the imbalance of all covariates altogether. For time dependent variables that are simultaneously confounders of the effect of interest can be estimated through Inverse-probability of treatment weighted (IPTW) methods (Wang and Fang, 2011). Propensity score matching of treated and control groups are very popular methods that reduce bias due to the covariates (Rosenbaum and Rubin, 1983). In contrast, propensity score matching has been challenged as a matching method that can increase imbalance if the propensity score model is mis-specified (Diamond and Sekhon, 2012). Another challenge is to assess the covariate imbalance across treatment groups i.e., when the distributions of relevant pre-treatment variables differ. A common approach that can reduce the imbalance between treated and control groups is Euclidean/Mahalanobis distance matching. One limitation of such distance metric is that if there is an extreme outlier in one covariate for a unit—the estimated variance for that covariate will be high, and Euclidean/Mahalanobis distance ignore the differences in that covariate. In a special case, Gu and Rosenbaum (1993) reported that if a binary covariate that takes values 1 and 0 with probabilities p and $1-p$; whenever $p \to 0$, Mahalanobis distance tries to match a rare treated unit with this covariate equal to 1. The performance of Mahalanobis distance matching in better in small data whereas the performance of propensity score matching better in large data set. Once the matched sample is selected through distance metric, very simple methods can be used to analyze the outcomes, and typical analysis of matched samples do not require the parametric assumptions of most regression methods (Rosenbaum and Rubin, 1985). Once the distance measure is determined–we look for the quantity of interest. The quantity of interest for the outcome analysis depends on the researcher objectives—for continuous response the most common estimand is average treatment effect (ATE) or average treatment effect on the treated (ATT) and odds ratio for the binary outcomes. Note, if a matching method that discards both treated and control units whose characteristics are dissimilar according to a pre-defined metric that do not result ATE or ATT. In this article, we focus on ATT to compare the performance of the estimation of largest caliper matching compare with other matching methods.

The study analyzes the effect of right heart catheterization (RHC) which play central role in identifying pulmonary hypertension, and reinvestigate whether RHC led to increase odds of severe clinical outcomes by several matching methods on Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT) data, previously analyzed by several authors (Connors et al., 1996; Hirano and Imbens, 2001). The data was collected to investigate the effectiveness of RHC in the initial care of critically ill patients. In this procedure, practitioners guide a special catheter (a small, hollow tube) called a pulmonary artery (PA)

catheter to the right side of patients' heart that measures pressures along the way. Practitioners believe that RHC help to reduce clinical outcome (death within 30 days of enrollment) whereas previous studies show that the clinical expense with right heart catheterization is high and could be potential risk to the ill patient. A detail description of the data are given in section 4. In this article, we review and reinvestigate the best practice for the use of RHC and estimate the ATT using different matching methods.

Section 2 describes matching methods that have been considered in this article. Section 3 describes a series of Monte Carlo simulations to examine the performance of these methods in estimating treatment effects. Particularly, we report on bias, standard deviation and mean square error (MSE) of the estimates. Section 4 presents analysis of the SUPPORT data. Finally, in Section 5, we summarize our findings.

2. Matching Methods

Several researches have been conducted to compare the matching methods. Elze et al. (2017) compared four propensity score matching methods to covariate adjustment on four cardiovascular observational studies. Austin (2014) compared 12 matching methods for 1:1 matching on the propensity score. Ming and Rosenbaum (2000) observed that substantially greater bias reduction is possible if the number of controls in match to each treated unit is not fixed. Gu and Rosenbaum (1993) compared optimal matching with nearest neighbor matching based on Mahalanobis distance. In this article, we considered six different matching methods: nearest neighbor matching with replacement (NNWR), nearest neighbor matching without replacement (NNWOR), optimal matching (OPT), full matching (FL), genetic matching (GM) and largest caliper matching (LC). The choice of selecting a matched sample differs in the methods and each serves to achieve specific objectives.

2.1 Nearest Neighbor Matching With Replacement

NNWR matching matches all treated subjects to their nearest control subjects based on a distance metric. In this method, each treatment subject can be matched to the closest control subject, even if that control subject is matched more than once. Because this approach can provide closer matches on the distance than nearest-available matching without replacement, it can be beneficial for reducing bias in the analysis. However, inference becomes more complex when matching with replacement, because the matched controls are no longer independent, some are in the matched sample more than once and this needs to be accounted for in the outcome analysis, for example by using frequency weights (Stuart, 2010). An illustration of the method is shown in Figure 1 using Mahalanobis distance metric to find the nearest control for the treated subjects.

2.2 Nearest Neighbor Matching Without Replacement

NNWOR requires that each match contains exactly one treated subject and exactly one control subject also known as 1:1 match. Once a control subject is matched with a nearest treated subject that control subject is no longer eligible for consideration as a match for other treated subjects. That is why, NNWOR is also known as "greedy" matching. There is a high

potential risk that a treated unit would be matched with a dissimilar control unit according to a pre-defined distance metric that preclude in estimating unbiased estimate. This method can be beneficial when there are enough good matches. The method is illustrated in Figure 2 using Mahalanobis distance metric to find nearest match.

2.3 Optimal Matching Without Replacement

The optimal matching method seek to match subjects to minimize a global discrepancy measure, like the sum of distances within matched sets (Rosenbaum, 1989). Greevy (2004) develops the idea to improve matching methods with the goal of optimizing the overall similarity of matched subjects. One can introduce simultaneous objective functions to optimize sum, variance and higher order moments as well (Zubizarreta et al., 2014). It does not make any difference if optimal matching with replacement compared to nearest neighbor with replacement. When there is intense competition for control units, optimal matching performs well. Figure 3 illustrates the method using exactly one treated subject with exactly one control subject that minimizes overall Mahalanobis distance.

2.4 Full Matching

Full matching considers that there exist at least one matched control (treated) subject for every treated (control) subject. Similarly, a fully matched sample consists of matched set that contains one treated unit and one or more controls (or one control unit and one or more treated units). One can choose $1 : k$ or $k : 1$ matching in full matching. The flexibility of this matching method can result in using more of the data at hand and yield more effective comparisons (in terms of effective sample size) and closest-possible matches on any given distance (Hansen, 2004). The choice of k depends on the optimization of the objective function. Figure 4 illustrates how the subjects would be matched using 1:3 full matching.

2.5 Genetic Matching

Diamond and Sekhon (2012) proposed genetic matching that automates the iterative process of checking and improving overall covariate balance to determine the given covariates' weight and ensures convergence to the optimal matched sample. They proposed a distance metric for the method that minimize the overall imbalance by minimizing the largest individual discrepancy based on p-values from paired t-test. Though the algorithm for genetic matching makes transparent certain issues, but the average run time of the algorithm is longer than other methods. Figure 5 presents the sample that would be matched using the method.

2.6 Largest Caliper Matching

A matching method can be affected by long edges as described in Figure 1, particularly the order in which subjects are selected for matching and the maximum permitted difference between matched subjects. The choice of long edges impact on the imbalance of covariates and the closeness of matching, and strengths the association between the confounding variable and the exposure to vary. We introduce a method that provides a heuristic approach to select the

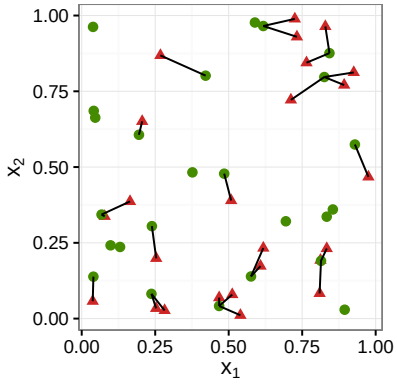

Figure 1:

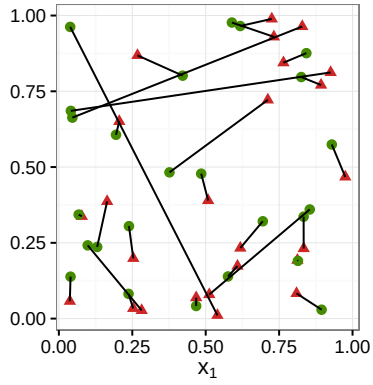

Figure 2:

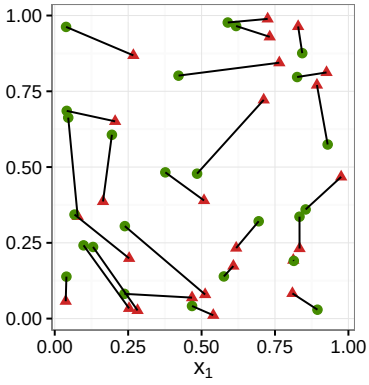

Figure 3:

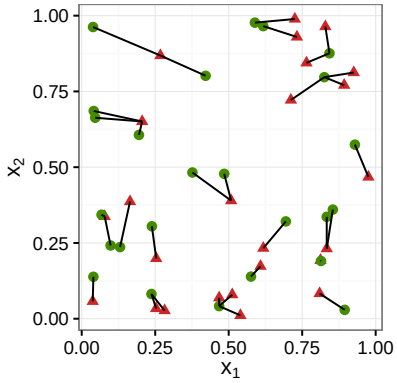

Figure 4:

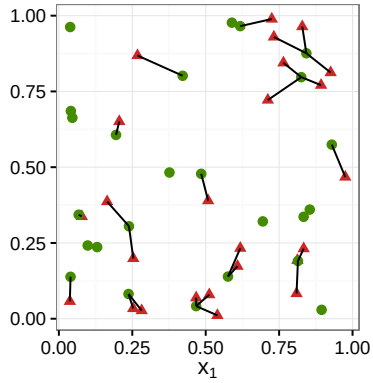

Figure 5:

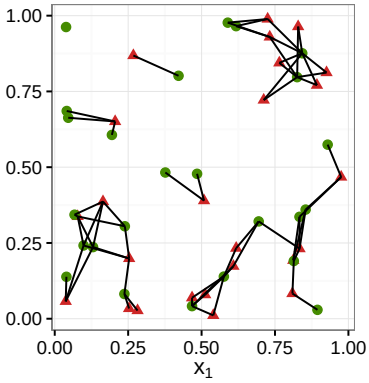

Figure 6:

Figure 7: Illustration of different matching methods. The sample consists of 50 subjects, both treated and control groups have 25 subjects each. We observe two covariates x_1 and x_2, for each subject. The red triangles indicate treated subjects and green circles indicate control subjects. Edges (based on Mahalanobis distance) indicate matched groups. A good matching method should avoid long edges, as they correspond to increase covariate imbalance.

maximum amount of imbalance that researchers want to accept for a match given a covariate, namely largest caliper matching (LCM). For largest caliper matching we consider the following distance metric:

The following distance metric is considered:

$$D^*(x_{ip}, x_{i'p}) = \max_p \frac{|x_{ip} - x_{i'p}|}{c_p}. \tag{2.1}$$

$D^*(x_{ip}, x_{i'p})$ is the amount of dissimilarity between ith treated and i'th control subject. Here c_p is a research-selected parameter for how much imbalance on covariate p is acceptable for a match. For example, if researchers want to match a treated subject of age 40 with a control subject of age within 35 and 45, then in this case $c_p = 5$. Similarly, if researchers want to match with a male treated subject with a female control subject then $c_p = 1$. A large value of c_p ensures a large number of matched subjects. For k categories, one can make $k - 1$ dummy variable and match in terms of the reference category or give weights to the units based on the

proportions of categories. If $D^*(x_{ip}, x_{i'p}) \leq 1$, then we say that an acceptable match. All the matched units that have at least one acceptable match as described in Figure 6—form a cluster of homogeneous subjects—are analyzed giving weights to the clusters based on the subjects in that cluster by total subjects.

We note several importance of largest caliper matching: First, LCM match based on the amount of imbalance that researchers want to accept for a covariate. For example, $D^*(x_{ip}, x_{i'p}) = 0$ means exact match based on pth covariate that researchers want to use for matching. Again, $D^*(x_{ip}, x_{i'p}) = \infty$ means match on pth covariate is negligible. Often it is not possible to reduce the imbalance for every covariate altogether, equation (2.1) might not be optimal by random choice of the c_p. We recommend choosing the c_p based on the important covariates that are related to the treatment assignment and study outcome. For large data set one can consider c_p as the caliper for the propensity score (Lunt, 2014). Austin (2011) observed optimal calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score when estimating differences in means and differences in proportions in observational studies. While the caliper matching (Austin, 2011) consider a caliper on propensity score—LCM consider a caliper both on propensity score and covariate space to discard the extreme observations. Second, LCM is a heuristic matching method—for a given c_p—the average run time of the method is faster than optimal matching. Third, the choice of c_p could be based on the quantity of interest. For example, if the quantity of interest is average treatment effect for the treated (ATT) (average treatment effect for the control (ATC)), then we chose the c_p in such a way so that every treated (control) subject has at least one matched control (treated) subject. Fourth, LCM is a version of cardinality matching, where within a given balance of the covariates, the maximum number of units that can be considered for analysis are considered. Fifth, LCM can be performed on any distance metric and often perform well on Mahalanobis distance metric for a small data. Finally, LCM ensures to discard the extreme units in the data that can increase the substantial bias in the analysis (King and Zeng, 2006) which is equivalent to give a negligible weight to that unit.

3. Monte Carlo Simulations

The simulations performed in the current paper employs a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results—simplistic matching simulations proposed in the literature (Austin, 2014; Pirracchio et al., 2015). We conduct a number of Monte Carlo simulations to compare the performance of six matching methods on binary outcome. Subjects were generated with $P(T = 1)$ so that effect of having $T = 1$ was assumed to increase a subject's probability of clinical outcome.

In each simulated sample, we compute an estimate $\hat{\tau}$ of the true parameter τ. We assessed the performance of each method using the following three criteria:

- Bias in estimating treatment effects: $\bar{\tau} - \tau$ where $\bar{\tau} = \sum_{l=1}^{N} \hat{\tau}/N$ and N is the total number of simulation.
- Standard deviation of the estimated treatment effect: $\sqrt{\sum_{l=1}^{N}(\hat{\tau} - \bar{\tau})^2/(N-1)}$.
- Mean square error of estimated treatment effects: $\sqrt{\sum_{l=1}^{N}(\hat{\tau} - \tau)^2/N}$.

3.1 The Setup

We considered $\boldsymbol{X}$ be a vector of 5 covariates that had effect both on the treatment assignment and the outcome. The treatment assignment model was generated from a linear combination of the covariates:

$$\text{logit}(\pi_t) = \beta_{0,t} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (\log(1.25), \log(1.5), \log(1.75), \log(2), \log(2))$. Thus, there were one covariate that had a weak effect on each of treatment effect and outcomes, one covariate had a moderate effect on each treatment assignment and outcomes, one covariate that had a strong effect on each of treatment assignment and outcomes, and two covariates that had a very strong effect on both treatment assignment and outcomes. The intercept of the treatment assignment model ($\beta_{0,t}$) was generated so that the proportion of subjects in the simulated sample that were treated was fixed at a desired proportion. We assigned treatment status (denoted by z) of subjects from a Bernoulli distribution with parameter π_t. The dichotomous outcome was generated using the following logistic model:

$$\text{logit}(\pi_o) = \beta_{0,o} + \tau z + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5.$$

We then generated a binary outcome for each subject from a Bernoulli distribution with parameter π_o. We selected the intercept, $\beta_{0,o}$, in the logistic outcome model so that the incidence of the outcome would be approximately 0.10 if all subjects in the population were control. In a given simulated data set, we simulated a binary outcome for each subject, under the assumption that all subjects were not treated ($z = 0$). We then calculated the incidence of the outcome in the simulated data set.

We selected the conditional log odds ratio τ so that average odds in treated subjects' due to treatment would be approximately 0.5. The same value of τ was used to generate a cohort of $n = 5000$ in a given scenario. Because we were simulating data with a desired ATT, the value of τ would depend on the proportion of subjects that were treated. This approach allows for variation in subject-specific treatment effects. The logistic model is used to simulate data with an underlying average treatment effect in the treated because such an approach will guarantee that individual probabilities of the occurrence of the outcome will lie within [0,1].

In Monte Carlo simulation, we consider a complete factorial design in which the following two factors were allowed to vary: (1) the distribution of the 5 pretreatment covariates; (2) the proportion of subjects that received the treatment. We considered four different distributions for the 5 pretreatment covariates: (i) the 5 covariates had independent standard normal distributions; (ii) the 5 covariates were from a multivariate normal distribution. Each variable had mean zero and unit variance, and the pair-wise correlation between variables was 0.25; (iii) the first two variables were independent Bernoulli random variables each with parameter 0.5, whereas the other three variables were independent standard normal random variables; (iv) the 5 random variables were independent Bernoulli random variables, each with parameter 0.5. For the second factor, we considered five different levels for the proportion of subjects that were treated: 0.1, 0.15, 0.2, 0.25, 0.3 and 0.35. Hence, there are 24 different scenarios of the study: four different distributions for the pretreatment covariates times six levels of the proportion of subjects that were treated.

In each of the 24 scenarios, we simulated $N = 1000$ datasets, each consisting of $n = 5000$ subjects. There were two reasons to use simulated datasets of size 5000. First, matching methods can be computationally intensive for large data. We considered a moderate size of the data that are available in real life. Second, researchers in different field usually have different size of the data—we observed in most cases these methods have been used in datasets of size around 5000. From the setup, we know the important covariates (i.e. x_4 and x_5) when matching and a good matching method should have more weight on these covariates. Though in real life

it is unknown that which variables are important for treatment and outcome but in practice—researchers use the existing literature or subject-matter knowledge and expertise to identify important variables that affect the treatment assignment or outcome. In each matched sample, we estimated the log odds ratio as the treatment effect. As the matched sample removes the effect of confounding due to pretreatment covariates—it was expected the estimates were unbiased.

3.2 Results

In Figure 11 we report the log odds ratio, standard deviation and mean square error of the log odds ratio when the pretreatment covariates were independently normally distributed. Figure 8 shows the bias of the methods under different treatment prevalence. A horizontal line has been added to each panel denoting the magnitude of the true log odds ratio 0.5. Figure 9 and 10 show the standard deviation and mean square error of the estimated log odds ratio, respectively. In general, as the prevalence of treatment increased the precision of the estimates increased for all matching methods. Optimal matching and nearest neighbor matching with/without replacement tended to have similar performance under independently normally distributed covariates. Amongst all methods, 1:3 full matching with caliper showed less standard deviation and mean square error of the estimated log odds. Largest caliper matching was the second choice in this scenario. Note that when the treatment prevalence was small, e.g. 10%, 1:1 nearest neighbor with/without replacement or optimal matching discarded at least 80% of the subjects from the data.

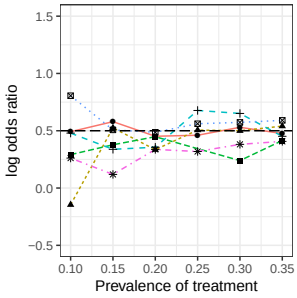

Figure 8:

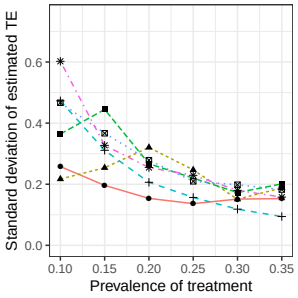

Figure 9:

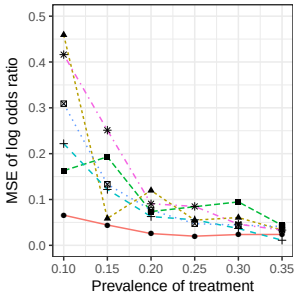

Figure 10:

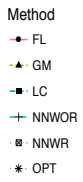

Figure 11: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under independent normally distributed covariates.

Figure 15 presents log odds ratio, standard deviation and mean square error of log odds ratio when the pretreatment covariates were multivariate normally distributed. The estimated treatment effect is reported in Figure 12. We see that nearest neighbor matching with replacement performs better than nearest neighbor matching without replacement. Largest caliper matching performed well through different treatment prevalence. The standard deviation and mean square error of the estimated log odds ratio are reported in Figure 13 and 14, respectively. Optimal matching and full matching showed less standard deviation and less mean square error in this case. The standard deviation was high for nearest neighbor matching with replacement when the treatment prevalence is low. Genetic matching performed better than any other methods when covariates were multivariate normally distributed.

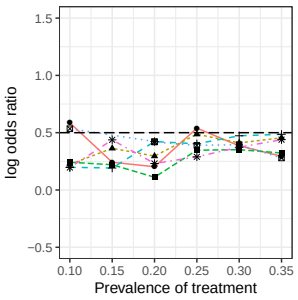

Figure 12:

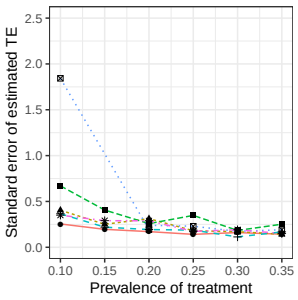

Figure 13:

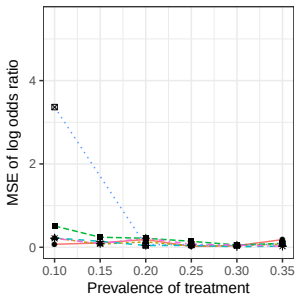

Figure 14:

Figure 15: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under multivariate normally distributed covariates.

In Figure 19 we report the log odds ratio, standard deviation and mean square error of the log odds ratio when the pretreatment covariates were both normally and binary distributed. Figure 16 shows the bias of the methods under different treatment prevalence. Largest caliper matching performed consistent over different prevalence of treatment. Figure 17 and 18 show the standard deviation and mean square error of the log odds ratio, respectively. Optimal matching and nearest neighbor matching with replacement had low precision in presence of low treatment prevalence. Both 1:3 full matching with calipers and largest caliper matching performed better than other matching methods.

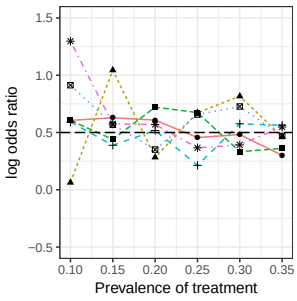

Figure 16:

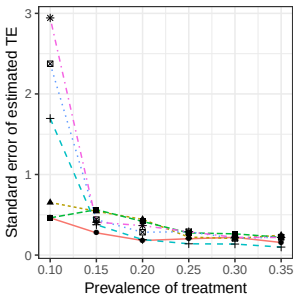

Figure 17:

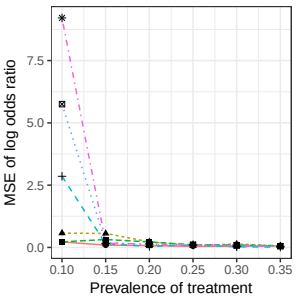

Figure 18:

Figure 19: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under both normally distributed and binary distributed covariates.

In Figure 23 we report the log odds ratio, standard deviation and mean square error of the log odds ratio when pretreatment covariates were independently binary distributed. Figure 20 shows the bias of the methods under different treatment prevalence. Both genetic matching and 1:3 full matching performed better than other methods in presence of low treatment prevalence. Figure 21 and 22 show the standard deviation and mean square error of the estimated log odds ratio, respectively. Nearest neighbor with replacement performed worse in this case.

In this section, we briefly discuss the results. In general, we observed several important facts that researchers need to consider in employing these matching methods. First, as the prevalence of the treated subjects increased from 10% to 35% in data, all methods tend to

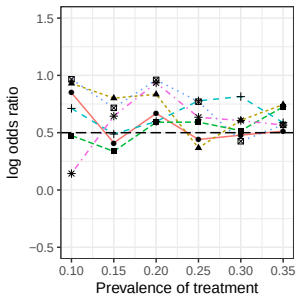

Figure 20:

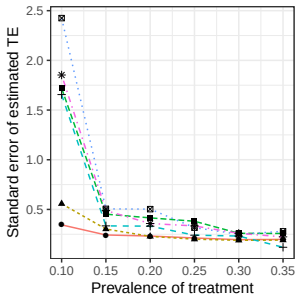

Figure 21:

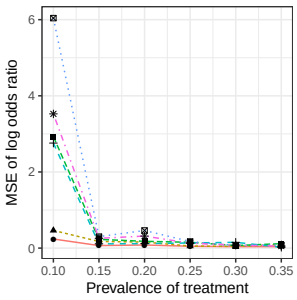

Figure 22:

Figure 23: Treatment effect: log odds ratio, standard deviation of estimated log odds ratio and mean squared error of log odds ratio under binary distributed covariates.

estimate unbiased estimate in the data and both standard deviation and mean square error of the estimates started to decrease. Second, full matching (in our case 1:3 with caliper) imposed more subjects than other methods—tended to result more precise estimates compared with the other matching methods. Note that full matching would perform better to reduce the covariate bias in the outcome analysis but could worsen covariate imbalance. Third, the choice between nearest neighbor with replacement and nearest neighbor matching without replacement reflected a bias-variance trade-off. In general, the nearest neighbor with replacement had lowest bias but higher variance compares to nearest neighbor without replacement. Some authors demonstrated this fact—matching with replacement produces matches of higher quality than matching without replacement by increasing the set of possible matches but have greater variability (Abadie and Imbens, 2006). Our result also support the following result through simulation and largest caliper matching takes into account both bias-variance tradeoff in the analysis. Fourth, when covariates have multivariate normally distributed covariates—genetic matching tended to have a performance that was at least as good as any of the competing methods. Fifth, we used Mahalanobis distance metric for nearest neighbor with replacement, nearest neighbor without replacement, optimal matching and full matching. In simulation we observed that for small number of covariates (in our case we considered five covariates) Mahalanobis distance metric performs much better than propensity score matching. Sixth, largest caliper matching considers an amount of covariate balance first then maximize the number of units within that balance whereas other methods iterate to improve the covariate balance. Though one can consider an optimal imbalance for largest caliper, it is recommended to use prespecified balance on important covariates only. Finally, our conclusions might be restricted to our simulation scenarios and might not apply to situations not represented by our simulated data. Again, a coverage of confidence interval might be another tool to check the performance of the methods.

The quantity of interest always depends on researcher objectives—that need to setup before analysis. If the number of control subjects are insufficient then nearest neighbor without replacement can result in exclusion of some treated subjects from the matched sample. Rosenbaum and Rubin (1985) used the term 'bias due to incomplete matching' to describe the bias that arises when treated subjects are excluded from the matched sample. In many real application, it is could be beneficial to discard some treated subjects without good match to obtain a good covariate balance. If a matching method discards treated subjects—the quantity of interest is no longer ATT. Since, in our simulation we considered the treatment prevalence maximum of 35%, our quantity of interest for all matching methods was ATT.

The results show that largest caliper matching performed fair under different setup. In

presence of large number covariates, we recommend to use all the covariates that are important for both treatment assignment and outcome. Unnecessary inclusion of covariates in the matching methods could reduce the performance of the methods (Stuart, 2010). Besides employing caliper on covariates—adding calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score for largest caliper matching in large data could make better performance. In this article, the analyses was conducted as a post-stratified sample—all the formed clusters were given weight to estimate ATT. In methodological literature, researchers have conducted substantial research on methods to estimate treatment effects. Besides, computationally they are very convenient—there are several `R` packages available for matching methods, e.g. `Matching`, `MatchIt` and `optmatch`.

We like to note certain attentions for the users of largest caliper matching. First, in LCM the analysis is sensitive to the choice of the caliper that could make substantial difference in matched sample. One choice of the caliper could be, consider only the important covariates that have higher standardized difference than a tolerance level. Second, a tighter caliper leads to reduce bias and make good matches but could discard those treated subjects that do not have good matches. Again, a flexible caliper increases the bias but reduces the variance of the estimate. Third, largest caliper matching ensures that there is at least one match for all treated subjects when the quantity of interest is ATT. Fourth, largest caliper matching is fast for a given amount of imbalance that researchers want to accept for a covariate. For SUPPORT data discussed in section 4, our largest caliper matching took 2.7 seconds to to run on a desktop computer with 2.7 GHz Intel Core i7 processor and 16.0 GB RAM. Fifth, largest caliper matching forms a good match sample that forms a cluster of homogeneous subjects. It successfully discards the control subjects that could increase the imbalance in the data. Finally, heterogeneity of treatment effects in subgroups of patients may provide useful information for the care of patients and for future research. Also, that influence the overall outcome of interest that is, while formed clusters have different treatment effects could be recognized by largest caliper matching. Combining these characteristics, largest caliper matching is a very computationally efficient and convenient matching method.

4. SUPPORT Data

The study analyzed SUPPORT data to investigate whether RHC led to increase odds of severe clinical outcomes, previously analyzed by several authors (Connors et al., 1996; Hirano and Imbens, 2001). SUPPORT data was collected on hospitalized adult patients at 5 medical centers in the U.S. Based on information from a panel of experts a rich set of variables relating to the decision to perform the RHC and outcome. Connors et al. (1996) found that after adjusting for ignorable treatment assignment conditional on a range of covariates, RHC appeared to lead to increase clinical death. This conclusion contradicted popular perception that RHC patients had less risk of clinical outcome. A detailed description of the study can be found in Connors et al. (1996) and Hirano and Imbens (2001).

The data had 5735 subjects, 2184 treated patients and 3551 control patients. There were 50 extensive covariates for covariate adjustment based on scientific knowledge, health status and clinical measures. For each subject, treatment status was observed equal to 1 if RHC was applied within 24 hours of admission, and 0 otherwise. Clinical outcome was an indicator for death within 30 days. There were (1486/2184) 68% of the RHC patients that had clinical outcome compared to (2236/3551) 63% of the No RHC patients. Fifty covariates were considered for covariate matching based on the covariates that are associated with the both RHC and clinical outcome. The presence of covariate imbalance is obvious in the data, for example,

there were (1278/2184) 59% Male in the RHC group compare to (1914/3551) 54% Male in the No RHC group. A detailed covariate imbalance picture has been given in Figure 24 by standardized difference. For continuous variables, the standardized difference is defined as $d = (\bar{x}_t - \bar{x}_c)/\sqrt{(s_t^2 + s_c^2)/2}$, where $\bar{x}_t$ and $\bar{x}_c$ denote the sample mean of the covariate in treated and control subjects, respectively, whereas s_t^2 and s_c^2 denote the sample variance of the covariate in RHC and No RHC groups, respectively. For dichotomous variables, the standardized differences are defined as $d = (\hat{p}_t - \hat{p}_c)/\sqrt{(\hat{p}_t(1-\hat{p}_t) + \hat{p}_c(1-\hat{p}_t c))/2}$, where $\hat{p}_t$ and $\hat{p}_c$ denote the prevalence or mean of the dichotomous variable in RHC and No RHC groups, respectively.

Apparently, the clinical measured variables (e.g., blood pressure, bilirubin level, etc.) are more confounded compare to the demographic variables (e.g., age, sex, etc.). In the propensity score model (40/910) 4% of the RHC subjects have lower (< 0.1) score and (87/97) 90% of the RHC subjects have higher propensity score. Out of 50 covariates there were 32 covariates that had absolute standardized differences were more than 0.1. NNWR and NNWOR matching data had 34 and 31 covariates that had absolute standardized differences more than 0.1. OPT performed better than nearest neighbor matching in terms of reducing covariate imbalance. LC successfully reduced all the covariate imbalances in the data and the result were consistent with other matching methods. Figure 24 reports the standardized difference for each of the 50 covariates in the matched and unmatched data.

Method	OR	2.5%	97.5%
NNWR	1.267	1.074	1.492
NNWOR	1.215	1.068	1.383
OPT	1.444	1.364	1.747
FULL	1.167	1.023	1.333
GM	1.243	1.097	1.409
LC	1.276	1.121	1.452

Table 1: Odds ratio of RHC group compare to No RHC group with 95% confidence interval.

We analyzed the unmatched data and the matched samples obtained from six matching methods. Table 1 shows the outcome analysis of the SUPPORT data. The second column presents the odds ratios of the analyses. We report that RHC was significant at 5% level of significance under all matching methods. This approach assesses and validate the inference to estimate the treatment effect. The interpretation of the analysis based on matched sample are more reliable compare to unmatched data.

5. Conclusion

The study proposes a new matching method and a review of the effectiveness of right heart catheterization (RHC) at a given time point, potential risk of RHC after adjusting other health factors. The study finds that survival chance is significantly worse for patients that undergo RHC after adjusting the factors that likely to influence the clinical outcome. Analysis of outcome in the unmatched data cannot adjust the confounders in a simple regression and may provide biased estimate of the treatment effect. Matching methods used to estimate the effect of interest, accounting for this confounding, to provide reliable estimate of the effect. This article presents more appropriate approaches for analyzing SUPPORT data. Our result further strengthens the substantive conclusion in Connors et al. (1996) that RHC lead to increased odds of severe clinical outcomes.

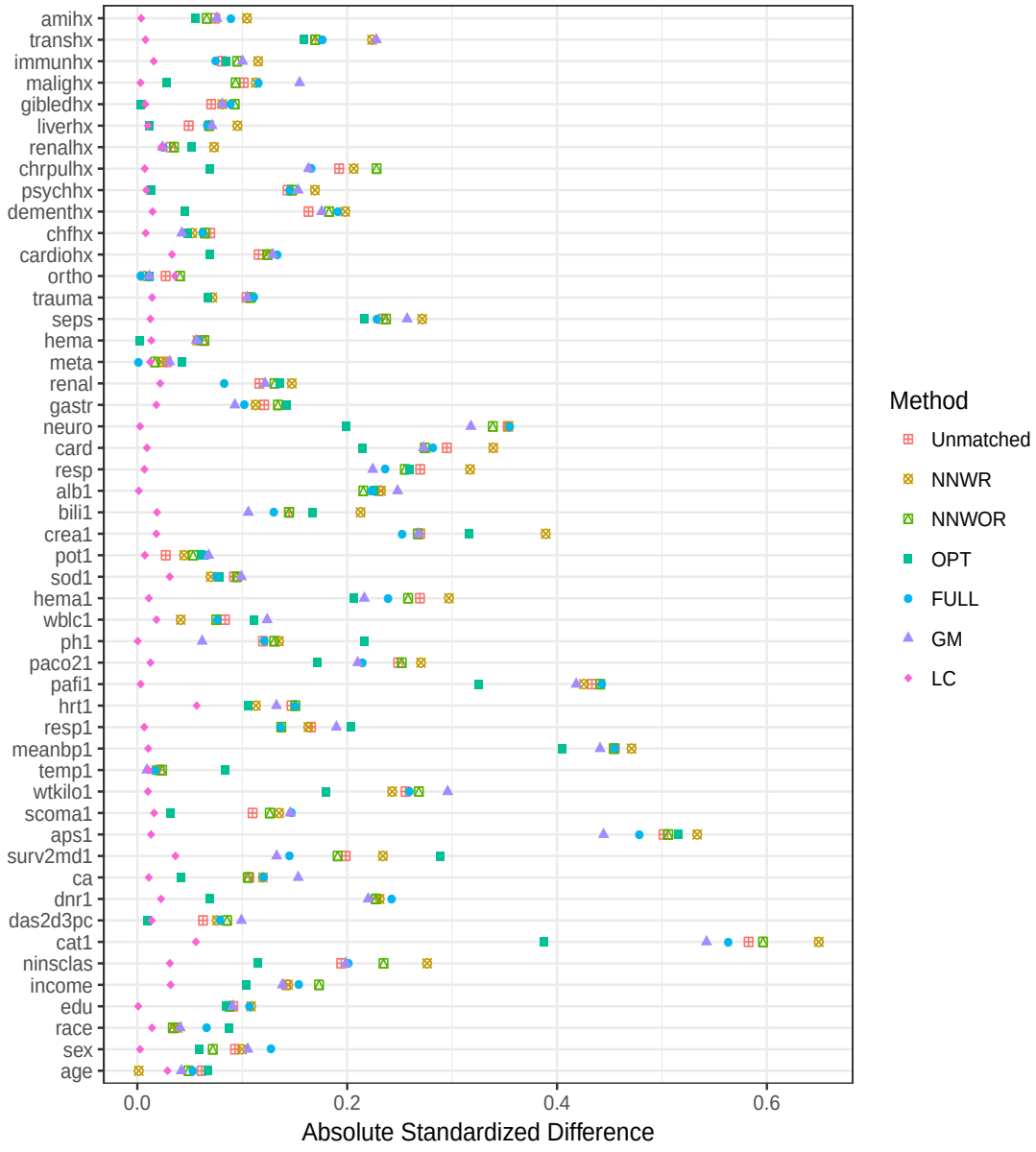

Figure 24: Covariate imbalance between treated/control subjects. The dotplot (a Love plot) shows the absolute standardized differences for unmatched and six matched samples.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, **74**(1):235–267.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, **10**(2):150–161.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, **33**(6):1057–1069.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., and Califf, R. M. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *JAMA : the journal of the American Medical Association*, **276**:889–897.

Diamond, A. and Sekhon, J. S. (2012). Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, **95**(3):932–945.

Elze, M. C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G. W., and Pocock, S. J. (2017). Comparison of propensity score methods and covariate adjustment. *Journal of the American College of Cardiology*, **69**(3):345–357.

Greevy, R. (2004). Optimal multivariate matching before randomization. *Biostatistics*, **5**(2):263–275.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, **2**(4):405–420.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, **99**(467):609–618.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, **2**(3):259–278.

Kang, J., Su X., and Liu, K. (2012). Tree-Structured Assessment of Causal Odds Ratio with Large Observational Study Data Sets. *Journal of Data Science*, **10**:757–776.

King, G. and Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, **14**:131–159.

Lunt, M. (2014). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American Journal of Epidemiology*, **179**(2):226–235.

Ming, K. and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, **56**(1):118–124.

Pirracchio, R., Petersen, M. L., and van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, **181**(2):108–119.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**:41–55.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, **84**(408):1024–1032.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**(1):33–38.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, **25**(1):1–21.

Wang, Y. and Fang, Y. (2011). Adjusting for Treatment Effect when Estimating or Testing Genetic Effect is of Main Interest. *Journal of Data Science*, **9**:127–138.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *The Annals of Applied Statistics*, **8**(1):204–231.