Match Making: A simulation study comparing the performance of 3 patient matching methods commonly used for causal inference from observational data

Jansen Storbacka, Laura Ruth

Match Making

A simulation study comparing the performance of 3 patient matching methods commonly used for causal inference from observational data

Laura Ruth Jansén-Storbacka

Thesis advisor: Professor Saskia le Cessie, Leiden University Medical Center
Thesis advisor: Professor Rolf Groenwold, Leiden University Medical Center

Defended on 16 August , 2022

MASTER'S THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Universiteit
Leiden
The Netherlands

Abstract

Matching methods are often used by researchers to improve causal inferences made from observational data. They aim to reduce confounding by improving covariate balance between treatment groups. Allowing more accurate estimates of causal effects. This research compares three commonly used matching methods; Nearest neighbour matching with replacement and without replacement, and propensity score matching using a caliper (PSM). Monte Carlo simulations were performed to assess the strengths and weaknesses of the methods in estimating the true treatment effect under different conditions. Two sources of bias were identified. The first occurs when there is incomplete overlap of the control and treated groups, even when there is theoretical positivity. The second is introduced when there are interactions between the outcome and the covariates. The two sources of bias combine and can either exaggerate or mitigate the total bias. The results show that when bias from overlap can be mostly removed by using PSM, this increases the number of treated units discarded. This means that only the average treatment effect on the matched (ATM) can be calculated, not the usual target estimand the average treatment effect on treated (ATT).

For higher sample sizes and a 1:1 ratio of controls to treated units, matching with replacement performed better than PSM. Using matching with replacement also removed a lot of the overlap bias without discarding any treated units. It was found that when using PSM the caliper is sample size dependent; a fixed caliper on the logit scale varies in size on the original covariate scale especially for low sample sizes. It was also found that the ideal caliper size, especially for smaller sample sizes and low levels of overlap can be between 0.3 *and* 0.7*, higher than the* 0.2 *that is commonly recommended. The results show how even a small lack of positivity in the form of overlap of* $< 100\%$ *can introduce bias. They also highlight the importance of interaction as an additional source of bias when in combination with lack of positivity. Additionally, this study demonstrates the need to pay attention to the proportions of discarded units when using PSM or matching with replacement as this affects the variance.*

Acknowledgements

I would like to express my gratitude to my supervisors Saskia le Cessie and Rolf Groenwold for their patience and support in guiding me through this project. Thank you both for your patience, in answering my many questions, reading many drafts, and helping me to understand unexpected results.

I would also like to thank my family for encouraging me to begin this and giving me the time in which to do it.

Contents

1 Introduction

Throughout history, people have attempted to understand the world by ascribing causes to outcomes. However, while this can lead to many useful discoveries, failing to separate correlation from causation however lead to many superstitions as well. Correctly pairing a cause with its effect is extremely important in medicine, where there are many maladies and many proposed cures. By randomly assigning treatments, randomized clinical trials allow the effects of treatments to be clearly identified, and separated from any effects that are merely due to correlation.

There are however many research questions where for ethical or practical reasons it is not possible to randomly assign treatments to populations, and the high standards of randomized clinical trials are not achievable. As an alternative, researchers often look to observational data such as patient records and population databases, with control and treatment groups selected from the database. While observational data sets are potential goldmines of information, the non random treatment assignment means there is a high risk of confounding. Confounding occurs when some covariates influence both the treatment allocation and the outcome, meaning causal effects cannot be accurately estimated.

In an observational dataset, the control and treatment groups often differ on certain covariates. For example, if older people are more likely to be assigned a blood pressure medication, then simply comparing the outcomes of treated and untreated individuals will be misleading and give a false estimate of the true treatment effect. In this case we can say that the groups are poorly matched or unbalanced on their covariates. Poor matching of control and treatment groups can also mean that the estimated treatment effect is potentially biased and can vary according to the method selected for analysis. This is known as model dependence (King & Nielsen, 2019), and is another way that confounding can lead to incorrect conclusions.

Matching methods can be viewed as a preprocessing technique for sampling from observational data (D. Ho, Imai, King, & Stuart, 2007). Matching processes match similar units from the control and treatment groups, creating a sub-sample where treated and control groups are balanced on all observed covariates. In this way a pseudo-randomized trial is created and causal effects can then be inferred. This study examines the most popular approaches to matching, and evaluates their effectiveness in both creating balanced groups and in correctly estimating the targeted treatment effect.

1.1 Causal Inference and the Potential Outcomes model

For both randomized and observational studies, causal inference involves estimating the effect of a treatment Z on an outcome Y. A treatment can be many things, for example a medication, a surgical

procedure, or a lifestyle choice. The outcome and how it is measured can also vary.

The potential outcomes model (Rubin, 1974) provides a conceptual framework for dealing with outcomes caused by a treatment. In this model, for each individual i there are two potential outcomes one for each of two categories of treatment assignment. If the individual receives the treatment, then $Z = 1$, otherwise if the individual does not receive treatment and is in the control group then $Z = 0$. Using this system, the outcome Y if an individual i receives treatment can be denoted as $Y_i(1)$ since $Z = 1$. Likewise, the outcome for the same individual i not receiving treatment can be written as $Y_i(0)$ since $Z = 0$. These two possibilities $Y_i(1)$ where individual i receives the treatment, and $Y_i(0)$ where they do not, are known as potential outcomes.

If an individual causal effect c exists then the expectation of the potential outcomes for the individual i will differ under the two treatments as shown in equation 1 (Hernán & Robins, 2018).

$$E\Big(Y_i(1)\Big) \neq E\Big(Y_i(0)\Big) \tag{1}$$

The magnitude of the causal effect c for an individual can be calculated as the difference between the two outcomes (equation 2), or alternatively as their ratio (equation 3).

$$c = Y_i(1) - Y_i(0) \tag{2}$$

or

$$c = Y_i(1)/Y_i(0) \tag{3}$$

The only issue is that for any individual i we can only ever know the outcome Y_i under one treatment condition. One outcome is observed, and the other only exists hypothetically; it is a result which did not happen, but which would have occurred if the treatment for individual i had been different. The event that did not occur is known as a counterfactual as it refers to a situation that is contrary to the facts. As we cannot observe what the value of Y_i would have been in the counterfactual situation, it is therefore impossible to observe the treatment effect for an individual. This impasse is known as the fundamental problem of causal inference (Holland, 1986).

Holland refers to two solutions to the problem of causal inference, the scientific and the statistical. In the scientific solution experiments are repeated and there is a large and unverifiable assumption that all units are homogeneous i.e., they do not differ from each other or change over time. While this may be reasonable in the physical sciences, this approach is less suited to clinical studies where these assumptions

are highly unlikely to be valid.

The statistical solution to the problem of causal inference uses the information about the population of treated and untreated individuals. Although the individual causal effects cannot be observed, under certain assumptions it is possible to estimate the average treatment effect in a population.

1.2 Assumptions for causal inference

In order to make inferences about causality it is necessary to satisfy certain assumptions.

1.2.1 The assumption of strongly ignorable treatment assignment

This is the assumption that conditional on the covariates X, the potential outcomes $Y(0)$ and $Y(1)$ are independent of the treatment assignment Z ((Rosenbaum & Rubin, 1983)). Equation 4 shows this assumption.

$$Y(0), Y(1) \perp\!\!\!\perp Z|X \tag{4}$$

Confounding is when the same covariates influence both the treatment group assignment and the outcome, making it difficult to disentangle causation. The assumption of strongly ignorable treatment assignment is effectively equal to the assumption of no unmeasured confounding. This assumption is met when both treated and control groups are balanced across all covariates. If the covariates are not balanced there can potentially be confounding, and preprocessing such as matching is necessary before an effect can be calculated.

1.2.2 The assumption of stable unit treatment value (SUTVA)

This assumes that an individual's treatment does not in any way affect the outcome of any other individual in the study (Stuart, 2010). In the current research it is assumed that this requirement is met .

1.2.3 The assumption of positivity

This is the requirement that at all levels of all covariates, the probability of a participant belonging in either the control or the treatment group is between 0 and 1. This means there must always be positive probabilities of an individual with any set of covariates belonging to the treated group and the control group. Positivity is assured in a randomized trial where the treatment assignment probabilities are controlled.

A formal definition of positivity is that given a binary treatment assignment Z, a vector of observed covariates X, and a probability density function $f(.)$, then if $f(X) \neq 0$, then $p(Z = z|X) > 0$ for all $z \in Z$ (Westreich & Cole, 2010).

Deterministic violations of the positivity assumption occur when it is impossible for individuals with certain covariate values to have one of the treatment conditions. In this case an effect cannot be inferred for individuals with those covariate values, but only for the range of covariate values where the positivity assumption is valid.

Random violations of the positivity assumption occur when the random nature of sampling causes some individuals to have a probability of being treated of 0 or 1 for that sample, even though this is not true in the wider population. In an observational setting there may be random violations of positivity, especially for smaller samples. In this case it is still possible to make inferences.

1.3 Estimands for treatment effects

To define exactly how to compare treatment groups in a particular study we determine the estimand. The estimand is the target, the quantity that we want to know. Two commonly used estimands are the average treatment effect (ATE) and the average effect of the treatment on the treated (ATT) (D. Ho et al., 2007).

The ATE is the difference between the mean of the treatment group μ_1 and the mean of the control group μ_2. Equation 5 shows how the ATE is calculated by subtracting the expected value of the control group from the expected value of the treatment group.

$$ATE = E(Y_i(1)) - E(Y_i(0)) = \frac{1}{n}\sum_{i=1}^{n} E[Y_i(1) - Y_i(0)|X_i] \tag{5}$$

In clinical settings the ATT (equation 5) is more commonly used as the interest is only in the effect on the treated group.

$$ATT = E[Y(1) - Y(0)|T = 1] = \frac{1}{\sum_{i=1}^{n} Z_i}\sum_{i=1}^{n} = Z_i E[Y_i(1) - Y \tag{6}$$

1.4 Estimating causal effects in randomized clinical trials

When the treatment effect c is the same for all individuals then the ATE and the ATT are equal. When the covariate distributions are the same between the two treatment groups, the groups are said to be balanced and both the ATE and the ATT can be estimated directly.

In randomized clinical trials, the assumptions of causal inference are assumed to be met as individuals are randomly assigned to either treatment or control groups before the treatment is assigned. This means that the control and treatment groups are only randomly different on both the observed and unobserved covariates. This meets the assumption of strong ignorability and ensures there is no confounding. Consequently, any differences between the outcomes of the groups can be attributed to the effect of the treatment.

This means that in randomized control trials the ATE and the ATT can both be estimated by treating the unknown values as missing data, and comparing the treatment groups.

1.5 Estimating causal effects in observational studies

As with a randomized clinical trial, the aim is to estimate the causal effect of a treatment by comparing treated and control groups.

However, when using observational data, the treatment is not randomly assigned, and the control and treatment groups are selected from the population. This means the assumption of strongly ignorable treatment assignment is not met, and the causal effect between the groups is difficult to estimate due to the potential for confounding. Confounding occurs when a covariate or confounder X influences both the treatment Z and the outcome Y, making it difficult to correctly estimate a causal effect of the treatment. (figure 1).

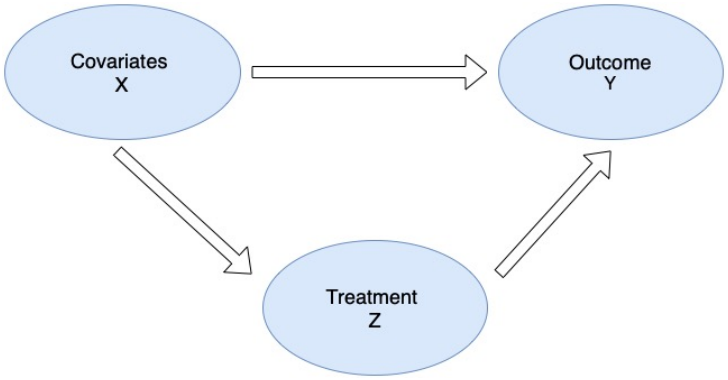


Figure 1: *Confounding occurs when covariates affect both the treatment and the outcome.*

Confounding can occur when the covariates are unbalanced because when the covariates are unevenly distributed, the influence of the covariates X on the treatment Z is different for each group, making certain covariate values more or less likely in the treated group.

1.6 Motivation for Matching

In order to estimate a treatment effect in observational data, the effects of possible confounding should be mitigated as much as possible.

Matching involves creating a more balanced dataset by selecting new treated and control groups from the original dataset. This is done before any analysis is performed on the data, and can be viewed as a way of simulating a randomized trial.

Patient matching can be useful in two situations (Stuart, 2010). The first is prospective, where a treated individual is selected from the general population, then followed and compared to a matched control individual. The second is retrospective, where all the data already exists, and the goal is to select the best matched set. In both these scenarios the researcher has no control over the allocation of treatment. The goal of matching is to allow more accurate inferences by improving balance and reducing confounding. As we can only match on the observed covariates however, causal inference from observational data is still inferior to inferences made using data from completely randomized designs.

1.7 An overview of Matching Methods

Creating balanced or matched samples can be achieved either by imputation of all missing counterfactuals as implemented in the R package Matching (Sekhon, 2008), or by subset selection. In this study matching refers to subset selection only.

Matching by subset selection involves sampling the treated and control groups in a way as to create a new subset which is balanced on the observed covariates. A range of subset selection methods are available within the R package MatchIt (D. E. Ho, Imai, King, & Stuart, 2011), comprising three main classes of matching algorithms; Stratum matching, Pure subset selection, and Distance Matching. All methods produce a new reduced dataset with improved inter-group covariate balance.

The pure subset selection method optimises to find the largest sample that still satisfies given balance and sample size constraints.

In stratum matching the multidimensional covariate space is divided on selected covariate values into zones called strata, and every stratum containing at least one of each treatment level is retained. Every individual i falls within a single stratum. Every stratum containing at least one of each treatment level $Y_i(0)$ and $Y_i(1)$ is retained and other strata are discarded. When there are more than one of either level within a remaining stratum then the units are weighted proportionally. In this way it is possible for one $Y_i(0)$ to be matched with many $Y_i(1)$ and vice versa.

When the strata are created using the original covariate values, this is called exact matching. Exact matching can work well when there are few covariates with either categorical or discrete numerical values, as it is feasible for matches to be made. With higher dimensional data however, the curse of dimensionality (Bellman, 1966) means that data become sparse and most strata contain zero or one

individual so a match cannot be found. A solution to this is to increase the size of the strata and use what is called coarsened exact matching (Iacus, King, & Porro, 2012). When there are more than one of either level within a remaining stratum then the units are weighted proportionally. In this way it is possible for one $Y_i(0)$ to be matched with many $Y_i(1)$ and vice versa. Another variety of coarsened exact matching is called subclass or coarsened propensity score matching, however this has fallen out of favour and been replaced by distance matching on the propensity score.

This study focuses on distance based matching methods as these are the most commonly used. Within distance matching the big questions considered are which distance measure to use, and whether to match with or without replacement.

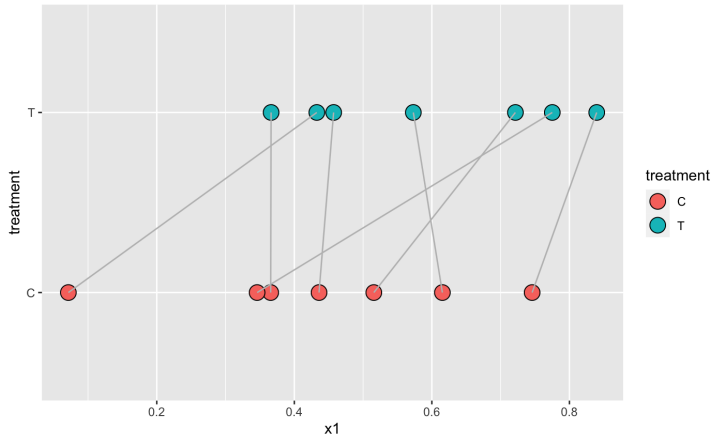
1.7.1 Distance Matching

In distance matching, matches are found by matching each treated individual to the closest control. First the type of distance measure must be chosen and then the distances between all pairs of treated and control individuals are measured. A common choice is the Mahalanobis or city block distance. Other possible distance measures include Euclidean distance, Cosine distance, and for categorical or discrete covariates Jaccard distance. The propensity score can also be used as a distance measure, this is discussed in the next section. The Mahalanobis and Euclidean distance measures are sensitive to scaling, this means that the units used for the covariates will affect the matches. This can be useful if it is the researcher wishes to give more important covariates greater influence in the matching process. Otherwise standardization should be performed to give all covariates equal weight. Cosine and Jaccard distances are scale invariant so they do not change with the units of the covariates and standardization is not required.

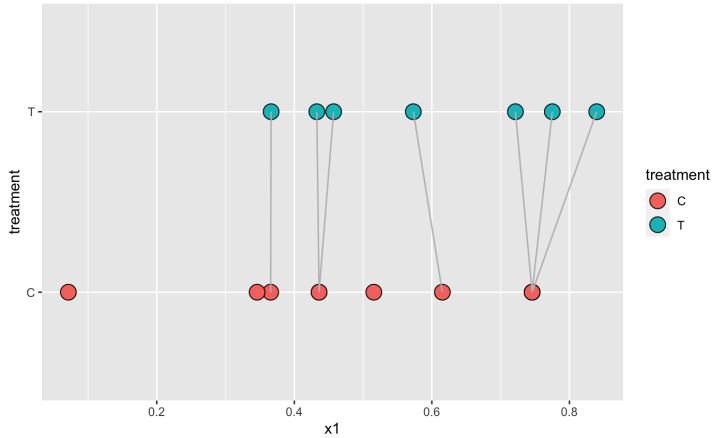
Distance based matching can be performed using $1:1$ matching, where every treated individual is matched to 1 control individual and the remaining controls are discarded. If as often occurs in population databases the dataset contains fewer treated than control individuals, it is also possible to use $1:k$ matching where each treated individual is matched to k controls.

The most common distance matching algorithm is nearest neighbour matching, this can be done using metric distances such as the Mahalanobis distance, or using propensity score matching. The nearest neighbour algorithm involves pairing a treated individual with the closest control individual. The distance is calculated for all possible combinations of confounders and the closest is selected. Nearest neighbour matching is a greedy algorithm. Greedy means that it chooses the best match for the current treated unit without consideration of how this will affect consequent matches or overall balance. This is especially important when there are few good matches available, and when the sampling is without replacement.

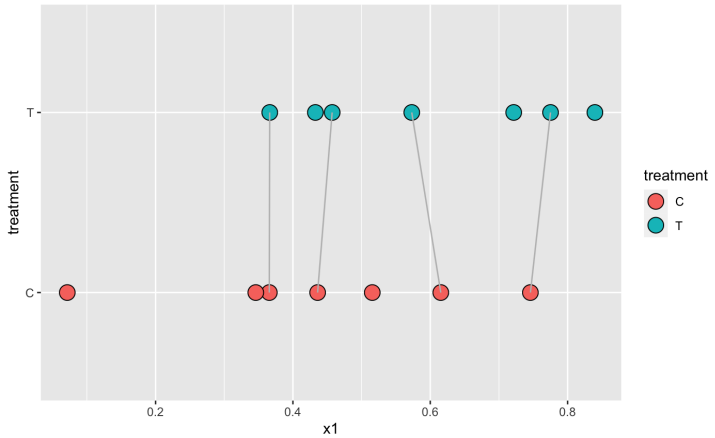
Optimal full matching and optimal pair matching are alternative types of distance matching. Instead of



(a) *Matching without replacement: Every treated unit is matched, however the greedy algorithm does not optimise overall so individual matches may be poor.*



(b) *Matching with replacement: Every control is matched to its nearest neighbour. This can lead to some controls being matched multiple times*



(c) *Propensity score matching without replacement and with a caliper: Only close matches are selected reducing overall bias. Both control and treated units can be discarded.*

Figure 2: *Examples of the matches made using three types of matching on a single dataset*

greedily making the best match for the current unit, optimal full matching optimizes the overall distances between all the matches. All units are segmented into sub classes such that each unit will have at least one match . Optimal pair matching is a sub category of optimal full matching.

Calipers only permit matches falling within a given range of the propensity score. They are often used to restrict the distance between matches. When calipers are used it is possible that no matches are found for some treated units, which are then discarded. This reduces the sample size available for analysis, but also reduces the number of poor individual matches. In this way it improves bias at the expense of the variance. Propensity score matching using a caliper is demonstrated for a one dimensional dataset in figure 2c.

1.7.2 Sampling with or without replacement

An important decision is whether to sample matches with or without replacement. When sampling without replacement using the nearest neighbour algorithm, a control unit is removed from the pool after it has been matched to a treated unit. In this situation, the sampling order can affect the matches as treated units that are matched later have fewer possible matches and later matches may be inferior. Figure 2a shows an example of matching without.

On the other hand, sampling with replacement means every treated unit is matched to its closest control neighbour, and no control units are discarded. Consequently, on measures of balance such as the difference in means, matching with replacement is always equal or better than matching without replacement. However when matching with replacement, it is possible for multiple treated units to match to a single control, reducing the size of the control group. This is demonstrated in figure 2b and is related to what King calls the balance-sample size trade off (King, Lucas, & Nielsen, 2017).

1.7.3 Propensity score theory

Instead of using nearest neighbour matching on multidimensional confounders, distance matching can be performed on a composite one dimensional propensity score. The propensity score $\pi(X)$ represents the probability for any individual of being in the treated group given its covariates (equation 7).

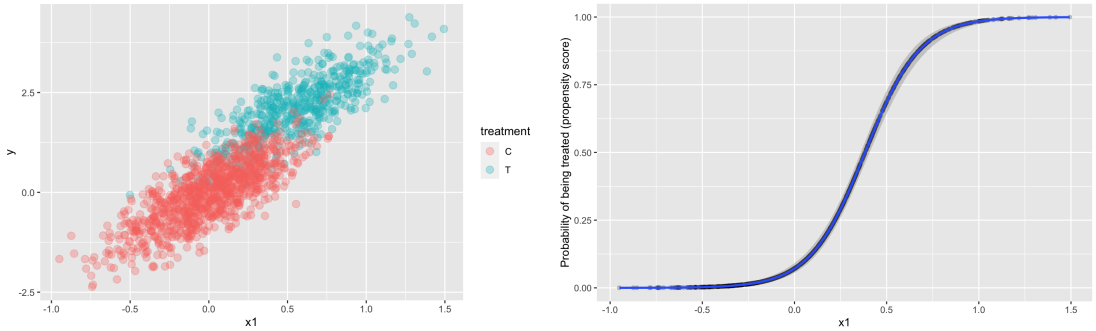
$$\pi(X) = Pr(Z = 1|X) \tag{7}$$

The true propensity score possesses what is called a balancing property (Rosenbaum & Rubin, 1983). This property means that for individuals with the same propensity score, the observed confounders will be balanced across the two treatment levels of the sample, even if individual matches are not optimal. This means that the sample will have improved overall balance of the observed confounders, and that

an average treatment effect for the population can be calculated. The theory behind propensity score matching states that for a given propensity score, matched pairs of treated and control units can be randomly sampled from the population. The mean difference of the treatment effects calculated between the matched pairs will then give an unbiased estimate of the true treatment effect.

1.7.4 Propensity Score matching

In practice, the true propensity score is unknown, and propensity scores are estimated using propensity score methods. Logistic regression is commonly used to estimate propensity scores by modelling the log-odds of $T = 1|X$. In figure 3a one dimensional, normally distributed treated and control groups are shown. Figure 3b shows the propensity scores at each value of the covariate for the data in (a), calculated using logistic regressionx_1 . Other classifiers such as decision trees can also be used to calculate propensity scores.



(a) *Normally distributed treated and control groups* **(b)** *Propensity scores for normally distributed data*

Figure 3: *Plot (a): Normally distributed treatment and control groups each with sample size* $n = 500$ *and standard deviation* $sd = \sqrt{\frac{1}{12}}$*.Means* $\mu C = 0$ *and* $\mu T = 2sd$*. Plot (b):Distribution of propensity scores for each covariate for the data in plot (a).*

1.7.5 Criticisms of Propensity Score Matching

Recently criticisms of propensity score have arisen (King & Nielsen, 2019). The main arguments are that (1) at best it is inferior to paired distance matching (section 1.7.1) as it has higher variance due to more units being discarded. (2) the theoretical assumption of balanced covariates does not always hold in practice, and individuals matched on propensity score may not be similar on their covariates, and (3) poor pruning where individuals are removed during the matching process can sometimes lead to increased imbalance. Coarsened exact matching where larger partitions between categories are used before exact matching is performed has been proposed as superior to propensity score matching (Iacus et al., 2012).

1.7.6 Assessing balance

To evaluate the quality of the matches found in observational data, it is necessary to measure the balance of the covariate distributions of the groups. The question of how control and treatment covariate distributions differ and whether or not they can be considered 'balanced' can be answered in various ways (Greifer, 2022).

The simplest and most commonly used balance measure is the standardized mean difference (SMD) between the covariate values of the two groups. For each covariate this is the difference between the means of the groups, standardized by dividing by the standard deviation of the treatment group. The closer the SMD is to zero, the better the matching is. As well as being simple to understand and calculate, the absolute standardized mean difference has been shown to perform as well or better in many situations (Ali et al., 2014).

Another method of assessing balance is the variance ratio which compares the spread of the distributions by comparing the variances of the control and treated groups. Similar distributions will have similar variances, giving a variance ratio close to one.

Empirical cumulative distribution function statistics (eCDF) for each covariate can also be used to compare distributions and assess balance. The Kolmogorov-Smirnov statistic can also be used to quantify the difference between distributions.

Visual assessment of the distributions is also an important way of assessing if the groups are balanced. This can be done using love plots, eCDF plots, emperical quantile-quantile (eQQ) plots, and kernel density plots

When simulated data is used the true treatment effect is known and it is therefore possible to calculate the bias of the estimate using Monte Carlo simulations. For this reason the balance of the matched groups is not examined in this study. However in real world studies the true effect is never known so approximating balanced groups allows for the best approximation of this.

1.8 Aims of the study

This study aims to examine the performances of three commonly used distance based matching methods. In particular it will examine nearest neighbour matching with and without replacement, as well as propensity score matching using a caliper to restrict matches.

The performance of each of these methods will be compared under different conditions. Simulated data will be used so that the true treatment effect can be used to measure performance. The aim is to provide

an overview of the strengths and weaknesses of the different methods and provide a road map for deciding how and if to match.

2 Methods

Monte Carlo simulations were performed to compare the quality of matches made using three common matching methods. In addition to the methods, six parameters were systematically changed in order to determine their effects. The quality of the matches was assessed by comparing the bias, variance and mean squared error for the different methods.

2.1 Simulation Aims

The three matching methods investigated were:

1. Nearest neighbour matching without replacement, with no caliper.
2. Nearest neighbour matching with replacement, with no caliper.
3. Propensity score matching without replacement, using a caliper.

The variable parameters used in the simulations were:

1. Sample size n, the number of treated units
2. The ratio of controls to treated units
3. The overlap or common support of the control and treatment groups
4. Caliper size (in the case of propensity score matching)
5. The size of the true treatment effect
6. The slopes of the treated and control groups; The effect of covariate value on outcome could be varied for the two groups. This allowed the creation of a heterogeneous treatment effect which changed according to the covariate value.

2.2 Overlap

Overlap refers to the values of the covariate included in both the treated group and control groups. This is also known as common support. Sufficient overlap is necessary for the assumption of positivity, as if there is too little overlap then not every unit has a positive probability of being in the treated or control groups, and the assumption is not met. In this study the overlap proportion was considered. This was defined as the proportion of the treated covariate values covered by the control covariate values.

2.3 Data Generating Mechanisms

To simplify the matching process, a single one dimensional covariate x was created for both treated and control groups. For each simulation a dataset was simulated representing units from both control and

treated groups. Random draws from a parametric model were used to generate a dataset of size n for the treated group and size $r \cdot n$ for the control group, with r representing the ration of controls to treated units. For each simulation either the normal or the uniform distribution was used for the covariate x, with both treated and controls being drawn from the same type of distribution, albeit with different parameters.

The uniform distribution was used because it provides a simple and evenly distributed dataset with clearly defined support. This enabled the proportion of common support between the groups to be calculated simply. The clearly defined support of the uniform distribution also means that it is easy to precisely manipulate the degree of overlap, and to push the boundaries of the positivity assumption.

The normal distribution was used firstly because it is a common distribution for many types of biological data. Secondly, as it is a continuous distribution and the support is asymptotically unlimited, so as n increases, even very different distributions will overlap and suitable matches can be found.

For the uniformly distributed data, the control group was distributed as

$x_C \sim U(a_C, b_C)$

and the treatment group as

$x_T \sim U(a_T, b_T)$

Likewise, for the normal distribution, the control group was defined as

$x_C \sim N(\mu_C, \sigma^2)$

and the treatment group as

$x_T \sim N(\mu_T, \sigma^2)$

The standard deviation of the normal distribution was chosen to be that of a uniform distribution $\sim U(0, 1)$.

this gave $\sigma = \sqrt{\frac{(B-A)^2}{12}} = \sqrt{\frac{1}{12}}$

This meant that the estimates would be more comparable between the experiments on different distributions.

An outcome variable was created according to the formula

$y = \beta_0 + \beta_1 x + \beta_2 T + \beta_3 xT + \varepsilon$

with $\varepsilon \sim N(0, 0.5)$

with T being the treatment group, β_0 the intercept, β_1 the slope of the control group, β_2 the treatment

effect, β_3 the interaction effect between group and x, ε the error term. The outcome variable was used to calculate the bias of the model made using the matched data, and was not used in the matching process.

This gave for each simulation a dataset consisting of three vectors, the treatment group T, the one dimensional covariate x, and the outcome Y.

The variable parameters used in the simulations with uniform data and their default values were;

- m the number of simulations performed (default $= 100$).
- n the number of treated units generated in each dataset (default $= 50$).
- r the ratio of control units to treated units generated (default $= 2$).
- The minimum and maximum treated values, Max T, and Min T (defaults $= (0.1, 1.0)$.
- The minimum and maximum control values, Max C and Min C (defaults $= (0, 0.9)$.
- The size of the true treatment effect at the intercept β_2 (default = 1).
- The standard deviation of the treatment effect (default = 0.5).
- The slope of the control group β_1 (default $= 2$ meaning the default $\beta_3 = 0$ and there is no interaction).
- The slope of the treatment group (default $= 2$, slopes of control and treatment can differ to introduce interaction effects).
- Whether to use replacement when matching control units (default = FALSE)
- Whether to use a caliper (default = FALSE)
- Which matching method to use (default = nearest)
- Distance (default = Mahalanobis)

2.4 Monte Carlo Simulations

In each simulation, A Monte Carlo simulation was performed, using the data generating process outlined above m times to create m unique samples (datasets) for each parameter combination. The same random seed was set before each simulation, to ensure any differences were due only to the methods compared.

2.5 Estimands

The target estimand was the Average Treatment effect on the Treated (ATT). This is possible to estimate when using methods such as nearest neighbour distance matching which use all treated values regardless

of the closeness of the matches. However when using a caliper in combination with propensity score matching some treated units may be discarded. When this happens is is no longer possible to estimate the ATT, instead the estimand that can be targeted is known as the Average Treatment Effect on the Matched (ATM). These simulations should help to determine when this difference in estimands is important.

2.6 Performance Measures

To measure the performance of the matching algorithms, it was necessary to consider not only the accuracy of the estimate, but also the reliability and whether the estimate was in fact measuring the target estimand.

2.6.1 Bias, variance and mean squared error

For each method the estimate was averaged over the m samples giving the mean of the estimated effects for both the ATT and ATM. Bias, variance and mean squared error were calculated for each method in order to compare the performance. Bias was calculated by subtracting the true treatment effect as specified when generating the data from the effect estimated from the matched data.

The variance of the estimates was also calculated and compared as it was also expected to be greater for those methods where units are discarded. When matching without replacement fewer control units can be used compared to matching without replacement. When matching using a caliper it is possible for treated units to be discarded, checking the variance ensures that the number discarded is not significantly altering the mean squared error. A similar situation applies with caliper matching as removing too many treated units from the matching process could increase the variance and subsequently the mean squared error.

The bias and variance were combined to give the mean squared error (MSE).

$MSE = \frac{1}{m}\sum_{n=1}^{n_{sim}}(\hat{\Psi} - \Psi)^2 = \text{Bias}^2 + \text{Variance}$

Where Ψ is the true value and $\hat{\Psi}$ is the estimate.

This was the most useful performance measure as it takes into account the accuracy as assessed by the bias, as well as adjusting for any effects on the variance caused by a reduction of sample size.

2.6.2 The proportion of unique control units when matching with replacement

For the matching with replacement method, one other measure was considered. As matching with replacement greedily chooses the best match every time using the shortest mahalonobis distance, it is possible for the number of unique control units to be lower than the number of unique treated units. This can mean that the variance is increased. It also means that the comparison is with a smaller number of

controls and as such is less representative of the general population. The bias is improved at the cost of the variance. For each matched sample the number of unique controls was divided by the number of treated. The proportion of unique controls was estimated by averaging this over m simulations.

$$\text{unique control proportion} = \frac{\text{number unique controls}}{\text{number of treated units}}$$

2.6.3 The proportion of treated units discarded when matching with a caliper

For matching using propensity score matching and a caliper the proportion of treated units discarded was also measured. This is important because it indicates how far the estimate, the ATM is from the target estimand the ATT. If a large proportion of treated have been removed then any causal inferences that can be made will be similarly limited.

2.7 Experiments

An overview of the experiments is provided in tables 1 and 2.

Table 1: *Overview of Monte Carlo Simulations assessing match quality for different matching methods. These simulations used randomly generated uniformly distributed data. Each situation was performed using m = 1000 repetitions, and the same random seed was used for each simulation.*

Experiments where x is uniformly distributed		
Experiment	**Simulations**	**Parameters Varied**
Effect of sample size	NN matching without replacement NN matching with replacement PSM + caliper	Sample size n = (1 - 100)
Effect of sample size ratio = 1, large sample	NN matching without replacement NN matching with replacement PSM + caliper 0.3 PSM + caliper 0.2	Sample size n = (50 - 1000)
Effect of Caliper Size and Overlap	PSM + caliper	Caliper size = (0.05 - 2) Overlap = (0.2 - 1.0)
Effect of sample size on caliper size on original scale	PSM + caliper	Sample size = (10,1010 caliper logarithmic scale = 0.3
Effect on treatment effect size	NN matching without replacement NN matching with replacement PSM + caliper	Treatment effect size (-2 ,+2)
Effect of Control:Treated ratio	NN matching without replacement NN matching with replacement PSM + caliper	Ratio of Control:Treated = (1 - 10)
Effect of varying overlap	NN matching without replacement NN matching with replacment PSM + caliper	Overlap proportion = (0 , 1)
Effect of interaction size (slope of treated)	NN matching without replacement NN matching with replacement PSM + caliper	Slope of treated = (1 , 6) Slope of control = 1
Effect of interaction as overlap changes	NN matching without replacement NN matching with replacement PSM + caliper	Overlap = (0,1) (for caliper, overlap = (0.35, 1.0)

Table 2: *Overview of Monte Carlo Simulations assessing match quality for different matching methods. These simulations used randomly generated normally distributed data. Each situation was performed using m = 1000 repetitions, and the same random seed was used for each simulation.*

Experiments where x is normally distributed		
Experiment	**Simulations**	**Parameters Varied**
Effect of sample size	NN matching without replacement NN matching with replacement PSM + caliper	Sample size n = (1 - 100) Far and close distributions
Effect of Caliper Size and Overlap	PSM + caliper	Caliper size = (0.05 - 2) Overlap = (0.2 - 1.0) Far and close distributions
Effect on treatment effect size	NN matching without replacement NN matching with replacement PSM + caliper	Treatment effect size (-2 ,+2) Far and close distributions
Effect of Control:Treated ratio	NN matching without replacement NN matching with replacement PSM + caliper	Ratio of Control:Treated = (1 - 10) Far and close distributions
Effect of varying overlap	NN matching without replacement NN matching with replacment PSM + caliper	Overlap proportion = (0 , 1) Far and close distributions

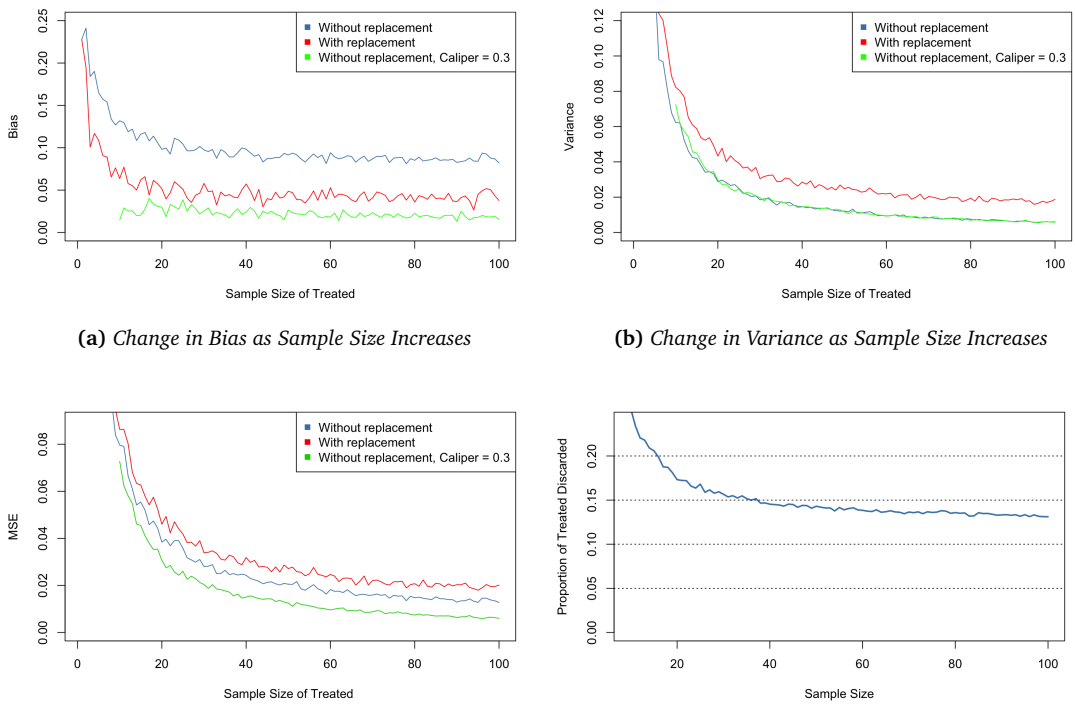
3 Results of experiments with uniformly distributed data

3.1 Default parameters

Uniform treated and control groups were generated with defaults of treated $X_T \sim U(0.1, 1)$ and Control $X_C \sim U(0, 0.9)$. This gave an overlap of 0.8, with the treatment extending beyond the common support of the control group by 0.1 and the control covariate values overlapping with around 89% of the treated covariate values. The treatment effect β_2 was set to 1, and the slopes of both treated and control groups β_1 were set to 2. The default ratio of controls to treated units was $2:1$.

3.2 Effects of sample size on match quality for uniformly distributed data

The default parameters were used here, and the sample size ranged from 1 to 100.



(a) *Change in Bias as Sample Size Increases*

(b) *Change in Variance as Sample Size Increases*

(c) *Change in MSE as Sample Size Increases*

(d) *Proportion of Treated Discarded using a Caliper*

Figure 4: *Changes in Bias, Variance, and Mean Squared Error (MSE) as sample size increases for three matching methods*

When increasing the sample size it can be seen that as sample size increases the bias of the estimates initially decreases and then plateaus for nearest neighbour matching with and without replacement. For PSM with a caliper, it is not possible to match with sample sizes less than 10. After this, the bias of the PSM estimates increases slightly around a sample size of 20 to 30 before decreasing again and plateauing. For all sample sizes bias is highest for estimates made by sampling without replacement. When sampling

without replacement it can be seen in figure 4(a) that the bias plateaus around 0.1. When sampling with replacement the bias is reduced, and it is lowest when using the propensity score with a caliper.

As expected the variance of the estimates decreases with sample size for all methods. Sampling with replacement gives the highest variance, with the other two methods performing similarly (fig 4b). This is because when matching with replacement variance is increased as there is a reduction in the number of unique controls used. This is examined more later (section 3.8). If the overlap were smaller, then there would be more treated units discarded and matches made using propensity score matching with a caliper would also have higher variance. In this scenario, however, the overlap is small and the number of discards is not high enough to greatly affect the variance.

For this uniformly distributed data with no interaction it appears that propensity score matching using a caliper gives the best result in terms of overall mean squared error (figure 4c. Sampling without replacement performs better than sampling with replacement due to the high variance contribution of the estimates when sampling with replacement.

It must be noted that propensity score matching using a caliper was unsuccessful when using smaller sizes $(n < 10)$ as the algorithm failed. In this case failure means that any one of the 1000 simulations failed to make a match. Looking at the number of discards (4d), it can be seen that for smaller sample sizes the average proportion of treated units discarded when $n = 10$ is close to 25%. It is important to remember that if any number of treated units are discarded then the estimate is no longer for the true ATT, and this is even more so with higher discarded proportions. For a genuine ATT, sampling without replacement appears to give the best results. Although it is more biased than sampling with replacement, the lower variance cancels this in the MSE.

Expected bias when matching without replacement

When matching without replacement, the expected bias can be simply calculated if the ratio is set to $1:1$. In this situation, the mean covariate value of the treated is $\frac{0.1+1}{2} = 0.55$ and the mean covariate value of the controls is $\frac{0+0.9}{2} = 0.45$. The expected value of the effect at each mean can then be determined.

$$\text{Expected outcome in treated} = \beta_0 + B_1 \cdot 0.55 + \text{true effect} = 0 + 2 \cdot 0.55 + 1 = 2.1$$

$$\text{Expected outcome in control group} = \beta_0 + B_1 \cdot 0.45 + \text{true effect} = 0 + 2 \cdot 0.45 + 1 = 1.9$$

The expected bias is then calculated by subtracting the mean control estimate from the mean treated

estimate, $2.1 - 1.9 = 0.2$.

If uniform data is simulated using a control to treated ratio of 1:1, and matched without replacement, it can be seen in figure 5a that the bias is 0.2. When matching without replacement and with a ratio of 2 controls to 1 treated as in this simulation, controls with covariate values less than 0.1 are never matched as a better match is always available for each treated unit. This means the mean of the available controls is 0.5 and the mean of the treated remains 0.55. This gives the expected outcome in the control group as $2 \cdot 0.5 + 1$. Subtracting from the expected outcome for the treated group gives $2.1 - 2 = 0.1$ which is what is found in the simulations (figure 4a).

Expected bias when matching with replacement

The expected bias can also be calculated when matching with replacement. As the sample size becomes larger, the treated units between 0.1 and 0.9 all find good matches, and their estimates will be unbiased. Only the treated units between 0.9 and 1.0 that are outside the common support of the control group will produce biased estimates. Of these the average value is 0.95 and for higher sample sizes this will always match to the maximum possible control value of ≈ 0.9. This means the matches for this portion of the treated outside the common support will on average be 0.5 higher than their matched controls. The expected bias when the slope of control and treated is 2 can then be calculated.

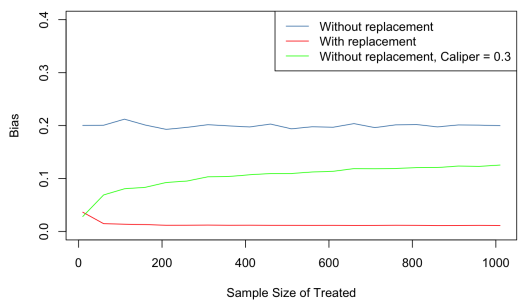
$$0 * 8/9 + 2 * 1/9 \cdot (0.95 - 0.90) = 0.011$$

3.3 Effect of sample size for higher n and when control to treated ratio = 1:1

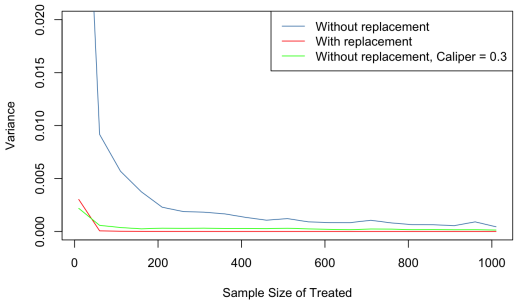
As discussed in section 3.2, the bias for matching data from the default uniform distributions without replacement stays close to 0.2 when the ratio of treated units to control units is $1:1$ (figure 5a), and this continues with higher sample sizes. The bias of the estimates for matching with replacement with a ratio of $1:1$ is also as predicted in section 3.2, as all estimates after $n = 300$ were 0.011 to 2 decimal places.

When matching using PSM and a caliper of 0.3, it can be seen in figure 5d that when the ratio of control to treated units is $1:1$, the bias increases slightly with the sample size. This increase is different from what was observed when matching with a ratio of $2:1$ and is most noticeable for lower sample sizes. This could be due to a combination of the high number of treated units discarded at lower sample sizes and the way the propensity score is calculated using logistic regression. This is discussed in greater detail in the discussion section.

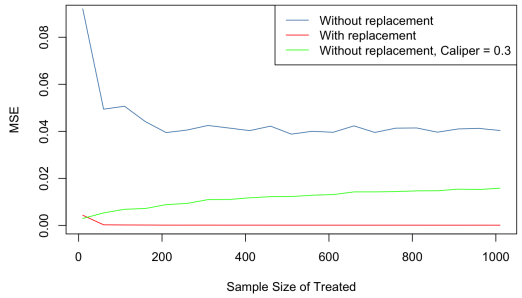
The estimates made by matching with and without replacement showed higher MSE for smaller sample sizes, while the estimates from PSM with a caliper showed a slight increase in MSE as the sample size



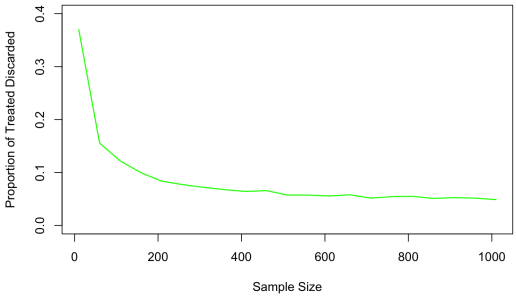
(a) *Bias of estimates*



(b) *Variance of estimates*



(c) *Mean Squared Error of estimates*



(d) *Proportion of treated units discarded (propensity score matching)*

Figure 5: *Bias, Variance, Mean Squared Error and Proportion of treated units discarded of estimates made using data uniformly distributed sampled with 3 matching methods, as sample size increases. Control to treated ratio = 1:1, overlap* $= 89\%$.

increased. For all but the smallest sample sizes, matching with replacement provided better quality matches in terms of both bias and MSE (figures 5a and 5c).

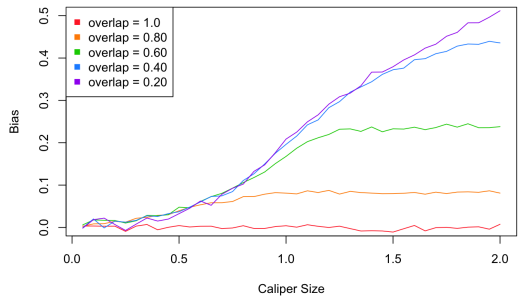
3.4 Effects of caliper size and overlap on match quality

When matching using the propensity score and a caliper, bias increases as the overlap proportion between the control and treated groups decreases. Bias also increases as the caliper size increases (figure 6a). This is because decreasing the overlap means fewer good matches are possible, and increasing the caliper means these less good matches are more likely to be accepted.

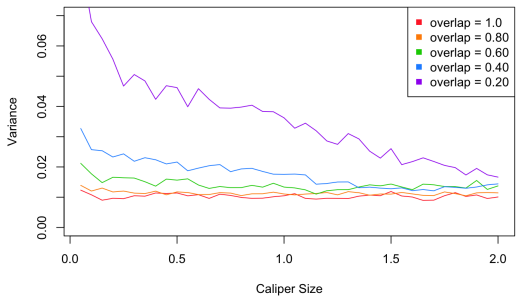
The variance decreases as caliper size and overlap increase (figure 6b). This is expected as increasing these also increases the effective sample size as fewer treated units are discarded.

Looking at figure 6c, it can be seen that the graph for the MSE is curved, reaching a minimum at caliper sizes between 0.3 and 0.7. This is due to the bias-variance trade off, as the high variance of the smaller caliper sizes removes the influence of the low bias. This is more important as overlap decreases.

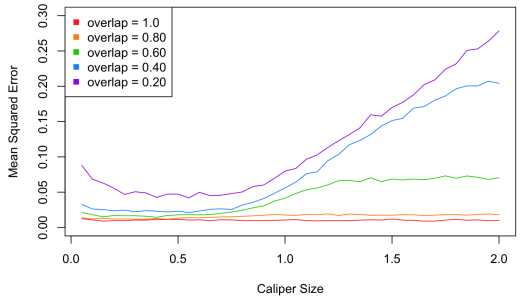
The proportion of treated units discarded decreases with caliper size, and increases for smaller overlap



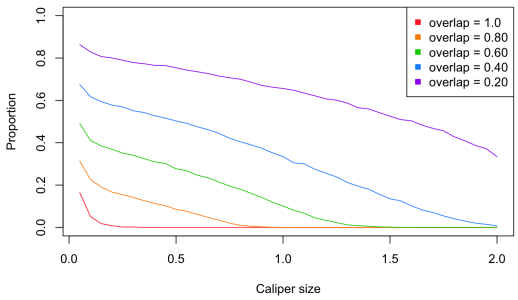
(a) *Change in Bias as Caliper Size and Overlap change*



(b) *Change in Variance as Caliper Size and Overlap Change*



(c) *Change in MSE as Caliper Size and Overlap Change*



(d) *Proportion of Treated Discarded by Caliper Size and Overlap*

Figure 6: *Changes in Bias, Variance, Mean Squared Error (MSE) and the Proportion of Treated Units discarded as Caliper size and Overlap (Common support) increase for uniform data. For overlap of* 20% *only* 500 *simulations were used as with* 1000 *the algorithm failed to converge.*

proportions (figure 6d). This is the cause of the changes in variance seen in figure 6b.

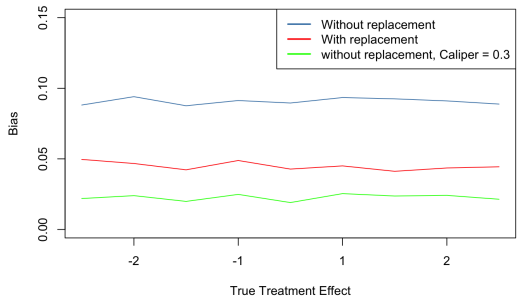
3.5 Treatment Effect Size and Match Quality

The size of the treatment effect does not influence either bias, variance or means squared error (7). This is not surprising, as the matching process is performed without knowledge of the outcome. Similar patterns to those previously seen are observed here.

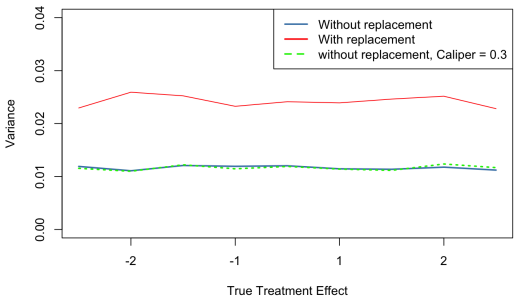
For the default high level of overlap and ratio of 2, we see that matching without replacement gives an estimate that is biased by approximately 0.09. This is slightly less than the bias of 0.1 that a ratio of $1:1$ would produce.

With the control to treated ratio of $2:1$, there are sufficient matches when using PSM, therefore the variance is equal to that with matching without replacement (figure 7b). As in the other simulations the variance is higher for the estimates made from matches without replacement. It can be seen from the jaggedness of the line that the estimate of the variance itself is more unstable.

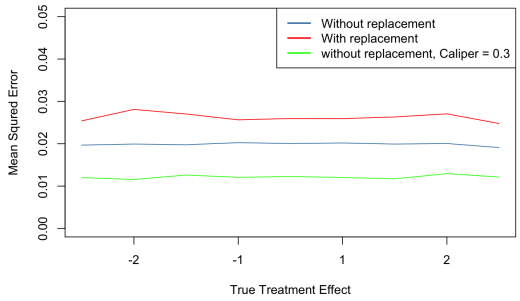
Looking at the MSE it can be seen that for this sample size and large overlap without interaction the PSM with caliper gave the most accurate estimates (figure 7c).



(a) *Bias as Treatment Effect Changes*



(b) *Variance as Treatment Effect Size Changes*



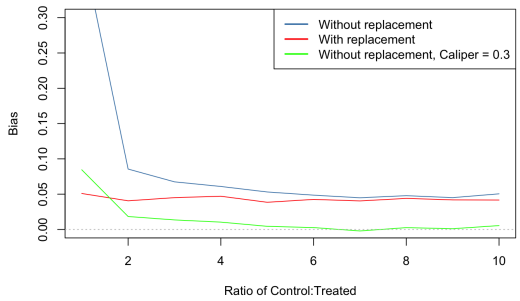
(c) *MSE as Treatment Effect Size Changes*

Figure 7: *Changes in Bias, Variance, and Mean Squared Error (MSE) as The True Treatment Effect increases for three matching methods*

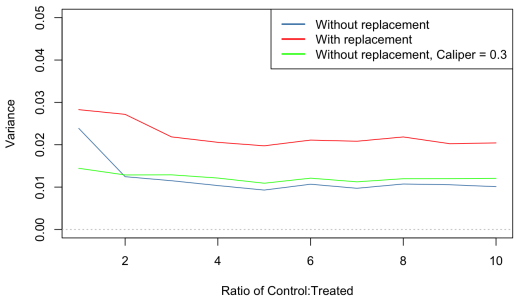
3.6 Effects of the control to treated ratio on match quality

When matching without replacement, the bias decreases sharply when increasing the control to treated ratio from $1:1$ to $1:2$. This is because matching with a $1;1$ ratio without replacement means all control units will be matched regardless of how bad the match is. When using PSM with a caliper, bias also decreases and this is most notable when changing from a ratio of $1:1$ to $2:1$. When the ratio is $1:1$, the bias for caliper matching (without replacement) and nearest neighbour matching (without replacement) is higher as there are often fewer suitable matches (figure 8a). When matching with replacement, increasing the ratio does not affect the bias.

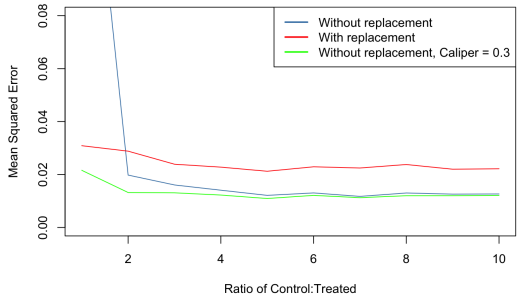
Variance also declines slightly for all methods (fig 8b). It decreases for estimates made by sampling without replacement as the algorithm is able to more consistently pick better matches leading to less variability. It decreases for matching with replacement as the number of unique controls increases. The decrease in variance for PSM using a caliper is very slight and is due to slightly fewer treated units being discarded as matches for the more extreme treated units are more likely to occur as the ratio of controls to treated units increases.



(a) *Bias as Control to Treated Ratio Changes*



(b) *Variance as Control to Treated Ratio Changes*



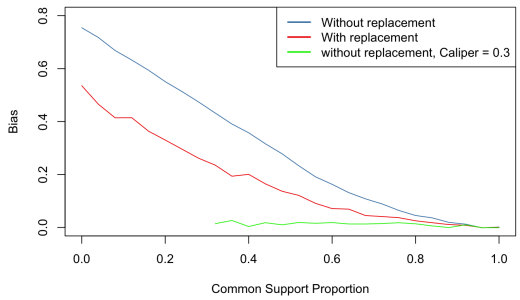
(c) *MSE as Control to Treated Ratio Changes*

Figure 8: *Changes in Bias, Variance, and Mean Squared Error (MSE) as the Control to Treated Ratio Increases for Three Matching Methods*

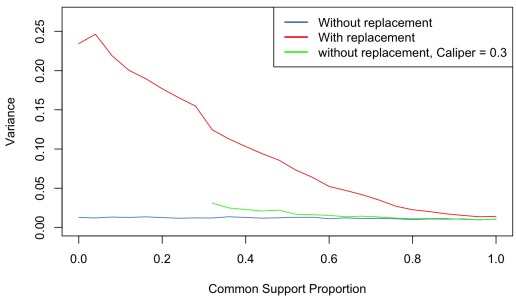
When matching with replacement the MSE only declines slightly as ratio increases (8c). For all methods the sharpest decline is from $1:1$ to $1:2$, and after this increasing the ratio has only very small effects on the overall error. This is consistent with the rule of thumb in clinical trials which says that a treated to control ratio greater than 1:4 provides negligible improvement in power.

3.7 Effects of changing the common support overlap of the control and treated groups on match quality

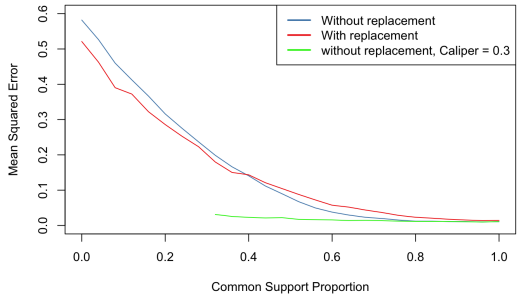
The bias of the estimates found using nearest neighbour matching with and without replacement declines as the common support proportion increases, and is higher for matching without replacement (figure 9a). Bias, variance and MSE are all lowest for matches made using a caliper (figure 9). However, with the default sample size of $n = 50$ and default ratio of $2:1$, if the common support proportion is lower than 0.35 it is not possible to match using a caliper of 0.3 as in some of the simulations there are no possible matches. For smaller overlaps sampling with replacement gives a lower MSE. For larger levels of overlap the MSE is smaller for sampling without replacement (while still estimating the ATT which the caliper matching does not). When the overlap is very high the methods perform similarly, as in this case the original data are already well balanced, and matching may not be necessary at all.



(a) *Bias as Common Support Proportion Changes*



(b) *Variance as Common Support Proportion Changes*



(c) *MSE as Common Support Ratio Changes*

Figure 9: *Changes in bias, variance, and mean squared error (MSE) as the overlap changes for three matching methods*

3.8 Effects of changing the control:treated ratio and overlap on the proportion of unique controls when matching with replacement

When sampling with replacement the proportion of unique values in the selected controls is affected by both the ratio and the overlap. The relationship between these parameters can be seen in table 3 and figure 10. In the simulations, treated and control groups were both generated using identical uniform distributions i.e. with a ratio of $1:1$ and 100% overlap. In this situation the proportion of unique controls was found to stabilize at around 0.56. This was also found to be true for different statistical distributions, as long as the distributions of the treated and control groups are identical. This finding was surprising as originally it was expected that the proportion would be similar to the bootstrap proportion of 0.63. Consequently, it was decided to calculate the expected number of unique controls, i.e the probability of being a nearest neighbour when the groups are of equal size and identically distributed (section 3.8).

Table 3: *The proportion of unique controls retained as overlap and ratio change*

Ratio C:T	Overlap 100%	Overlap 80%	Overlap 60%	Overlap 40%	Overlap 20%
1	0.562	0.451	0.340	0.227	0.115
2	0.721	0.584	0.438	0.293	0.150
3	0.798	0.644	0.485	0.324	0.166
4	0.843	0.679	0.511	0.344	0.176
5	0.868	0.703	0.531	0.356	0.183
6	0.891	0.718	0.540	0.364	0.187
7	0.905	0.732	0.550	0.371	0.190
8	0.916	0.737	0.556	0.374	0.190
9	0.923	0.748	0.563	0.378	0.196
10	0.931	0.754	0.568	0.383	0.197

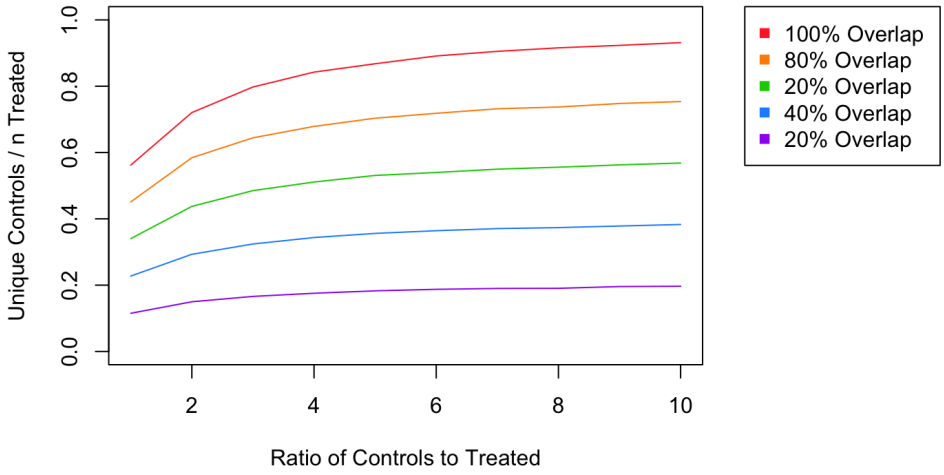


Figure 10: *The proportion of unique controls for different common support ratios and different control:treated ratios*

The expected proportion of unique controls when matching from 2 identical distributions with a control to treated ratio of $1:1$

For the control and treated groups we take two identical uniform distributions X_C and X_T both given by $\sim U(0,1)$ and draw n samples from each with $n \geq 2$.

First we calculate, for any point in the control group xc_i, the probability of being the nearest neighbour of a given point xt in the treated group, where $0.5 \leq xt \leq 1$.

When $xc_i = xt$ then the probability that xc is the nearest neighbour of xt is 1.

As xc is further away from xt, the probability of being a nearest neighbour declines. When $xc = 0$, then we can say that $p(xc = NN) = 0$, as since $xt \geq 0.5$ then another control point must be closer.

This gives the probability of being a nearest neighbour a slope of $\frac{1}{xt}$ between $0 and 1$. We assume the

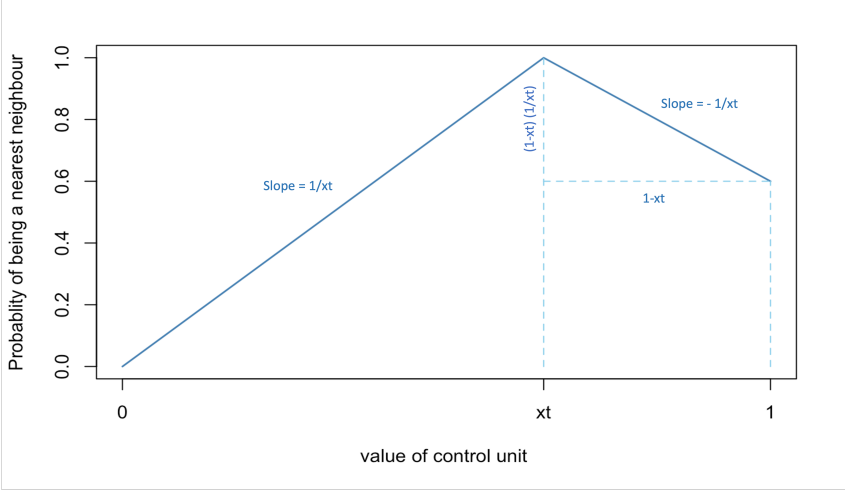


Figure 11: *The probability of any control X_C unit being a nearest neighbour of a treated unit xt where $xt > 0.5$ and is drawn from $X_T \sim U(0,1)$ and $X_C \sim U(0,1)$*

distribution is symmetrical at xt, then the slope between xt and 1 is $\frac{-1}{x}$, and $p(xc = NN)$. This means that when $xc = 1$, $p(xc = NN) = 1 - (1 - xt)(\frac{1}{xt})$.

This situation is shown visually in figure 11. The area under the curve (AUC) can then be calculated.

$$\begin{aligned} AUC &= \frac{1}{2}xt + (1-xt)(1-(1-xt)\frac{1}{xt}) + \frac{1}{2}(1-xt)((1-xt)\frac{1}{xt}) \\ &= 2 - xt - \frac{1}{2xt} \end{aligned}$$

This gives the probability for a fixed value of xt. Integrating this over all possible values of xt ($xt \sim U(0,1)$) gives the average probability of being selected as a nearest neighbour $p(NN)$

$$\begin{aligned} p(NN) &= 2 \cdot \int_{0.5}^{1} dxt \left(2 - xt - \frac{1}{2xt}\right) \\ &= 2\left[2xt - \frac{1}{2}xt^2 - \frac{1}{2}\log xt\right]_{0.5}^{1} \\ &= 2\left(\frac{5}{8} + \frac{1}{2}\log\frac{1}{2}\right) \\ &\approx 0.56 \end{aligned}$$

This expected value of 0.56 is the same as found in the simulations with a control to treated ratio of $1:1$ and overlap of 100% (table 3 and figure 10).

From this starting point, the number of unique controls rises consistently as the overlap proportion increases (figure 10). The proportion of unique controls also increases slowly as the ratio increases. If the ratio is sufficiently high the proportion of unique controls eventually reaches 100%.

3.9 Effects of interaction size on match quality

Bias when interaction is present

When the slopes of the measured effects of the treated and control groups differ, the size of the treatment effect depends on the value of the covariate. This is called an interaction effect. In an interaction scenario, there are two sources of bias, demonstrated here in figure 12. The first source of bias is caused by the overlap. As seen in the earlier simulations (figure 9a), overlap of less than 100% can lead to bias when treated points outside the area of common support make poor matches. As the slope of the effect on the treated increases, so does the bias caused by overlap. Here this gives a positive bias, much of which is removed when using PSM and a caliper.

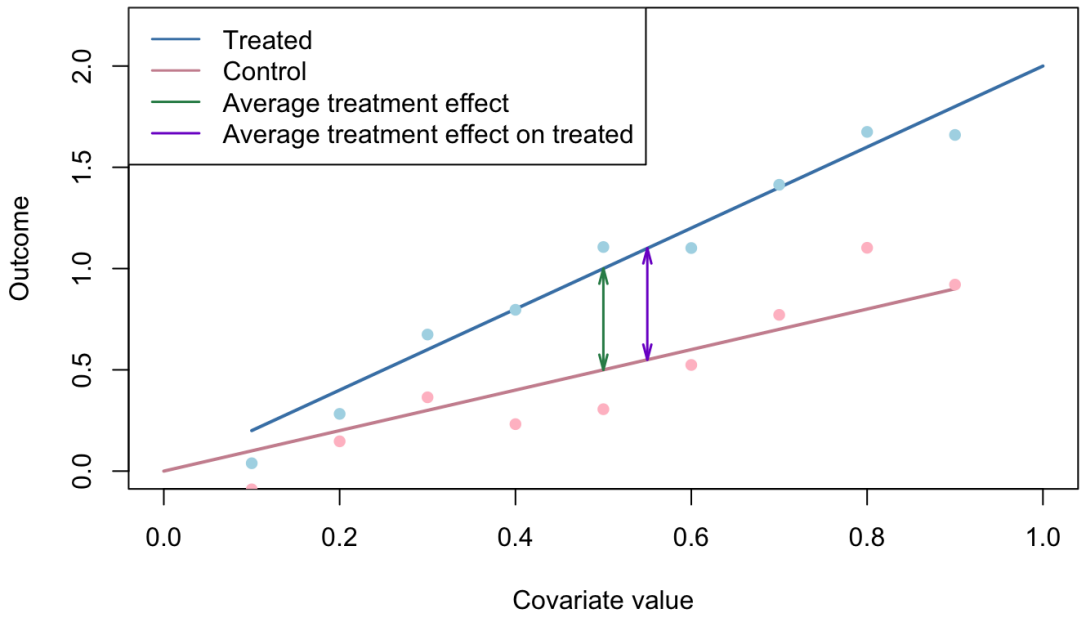


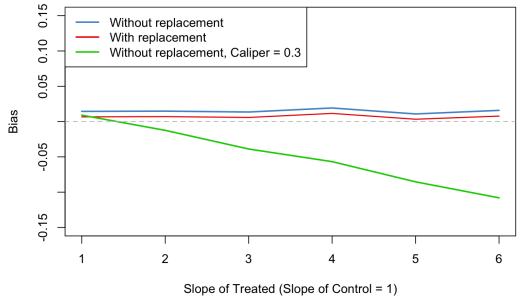
Figure 12: *The effects of interaction on estimated treatment effect where slope of control = 1 and slope of treated = 2*

The second source of bias is caused by the interaction. When interaction is present, $ATE \neq ATT$, however, the ATE is used to estimate the ATT and this is a cause of bias. In this example the bias is negative as the slope of the treated is greater than the slope of the control. Matching with PSM removes most of the treated values that occur beyond the maximum control value. The consequence is that when matching with PSM only the negative bias caused by the interaction remains.

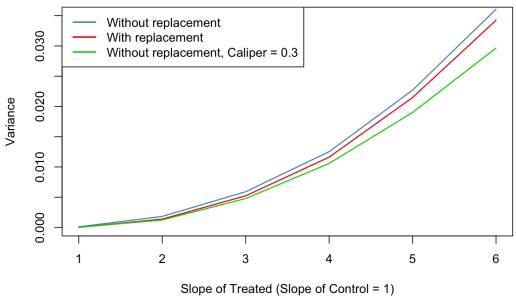
Simulation Results

In the simulation here the treated group is given by $X_T \sim U(0.1, 1.0)$ and the control by$X_C \sim (0, 0.9)$. The slope of the control was fixed at 1 and the slope of the treated was varied from 1 to 6. The ATE was calculated at the mean covariate value of 0.5, and the ATT at the treated mean of 0.55. The bias was calculated by subtracting the ATT from the ATE.

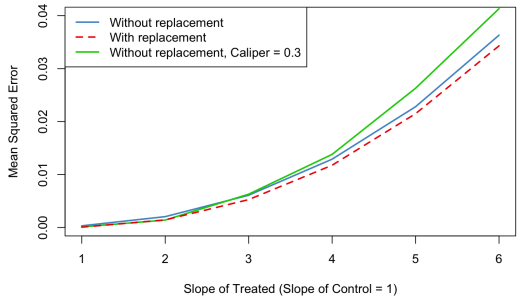
When matching using PSM and a caliper, removing extreme treated values removed most of the overlap bias meaning only the negative bias from the interaction remained. As the slope of the outcome of the treated group increases, the difference between the ATT and the ATE also increases. Consequently the negative bias becomes more pronounced (figure 13a).



(a) *Bias as interaction effect increases*



(b) *Variance as Interaction Effect Increases*



(c) *MSE as Interaction Effect Increases*

Figure 13: *Changes in Bias, Variance, and Mean Squared Error (MSE) as Interaction effect Changes for Three Matching Methods*

When matching with and without replacement the two opposing sources of bias mean the total bias is close to 0 (figure 13a). Matches made using sampling with replacement producing slightly less biased estimates than matches without replacement. When the overlap bias is removed by the caliper, the estimate of the ATE is used instead the ATT. When matching with or without replacement, the ATT is calculated more accurately and only the bias of the overlap is left.

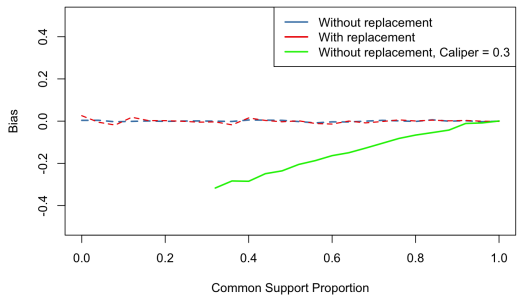
The variance of all matching methods also increased as the slope of the treated group increased (figure

13b). This is dues to poorer matches giving more unstable estimates.

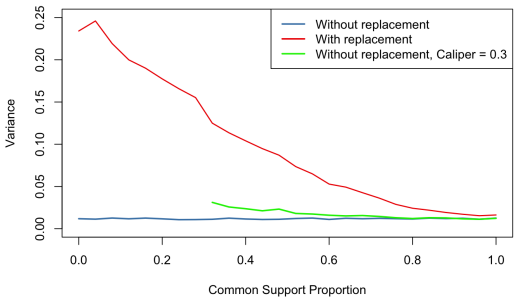
The MSE also increased with the size of the interaction effect. It increased the most for the PSM estimates, reflecting the contribution of the strong negative bias in this simulation.

3.10 The effect of varying overlap in the presence of interaction

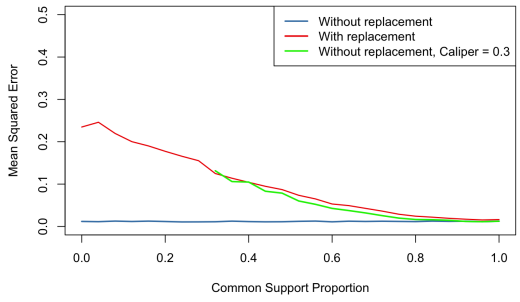
An interaction effect was created by fixing the slope of the control outcome to 0, and the slope of the treated outcome to 2. The overlap proportion was then varied from 0 to 1. The estimates for PSM with a caliper could not be calculated for overlap proportions less than 0.35, as too few suitable matches were found. When matching with and without replacement bias was low for all levels of overlap (figure 13a). When using a caliper, the negative bias caused by the interaction effect (subsection 3.9) dominated more as the overlap decreased. For low levels of overlap the matches made by sampling with replacement gives estimates with higher variance (13b). This indicates that number of unique controls selected is very low. For lower levels of overlap, the high variance of the estimates from sampling with replacement increase the MSE. On the other hand, for estimates made using PSM, higher bias contributes to higher levels of MSE (13c). In this situation, with both interaction and low levels of overlap, the best estimates were given by nearest neighbour matching without replacement.



(a) *Bias as Common Support Changes (with interaction)*



(b) *Variance as Common Support Changes (with interaction)*

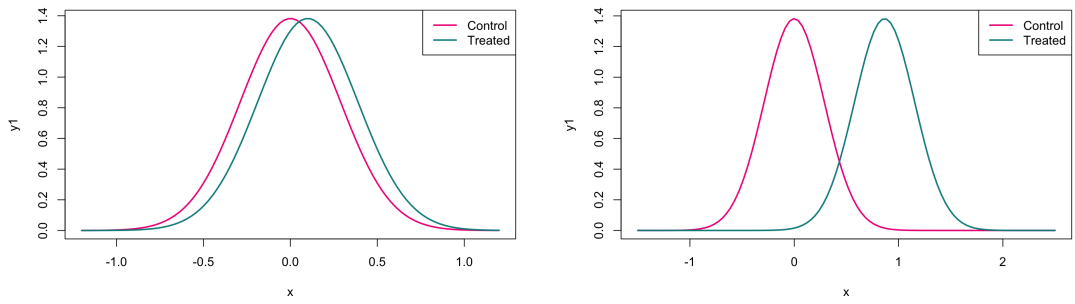


(c) *MSE as Interaction Effect Increases*

Figure 14: *Changes in Bias, Variance, and Mean Squared Error (MSE) for Three Matching Methods as Overlap changes, when interaction is present*

4 Results from experiments with normally distributed data

For the experiments with normal data two sets of normal distributions were used. The normal distribution experiments were performed using both close (figure 15a) and far (15b) distributions for the treated and control groups. The close dataset had high overlap of the control and treated distributions with only 0.1 difference between the group means. In the low dataset the distributions had 3 standard deviations difference between their means. For all distributions the standard deviation was set at $\sqrt{\frac{1}{12}}$ as this is equal to the standard deviation used in the experiments using the uniform distributions. The two levels of common support were chosen to compare the performance of the different methods under these different circumstances.



(a) *close normal distributions* **(b)** *far normal distributions*

Figure 15: *The two normal distributions used for the simulations both have standard deviations of standard deviation is* $\sqrt{\frac{1}{12}}$. *Figure a shows close treated and control distributions with a high level of overlap, the distance between the means is 0.1. Figure b shows normal treated and control distributions with low overlap, the distance between means is 3 times the standard deviation*

4.1 Effect of sample size on match quality for normal distributions with large and small overlap

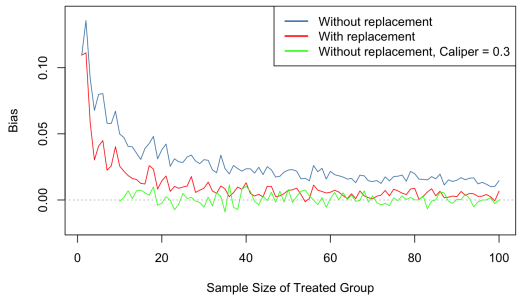
In figures 16a and 16b it can be seen that the bias follows a similar overall pattern to the bias in the uniform distribution (16b), as it decreases with sample size, and is highest for sampling without replacement. For the far normal distributions however, the overall pattern of biases is much more extreme, with all methods showing higher bias when compared to the close distributions. The without replacement group in figure 16b shows especially high bias and very little improvement with sample size. For the highly separated far distributions, propensity score matching with a caliper did not match with a sample size less than 40.

The variance also showed a similar pattern to the uniform experiments, with the same tendencies exaggerated more for the far distributions (figures 16c and 16d). For far data, the variance is much higher for the with replacement method, this is due to few controls being selected.

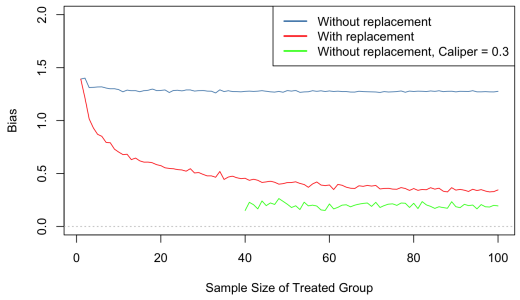
When bias and variance are combined in the MSE, the results differ for the close and far distributions

(figures 16e and 16f). For the close distributions the best overall estimates were given by the matches sampled without replacement, although all had very low MSE.

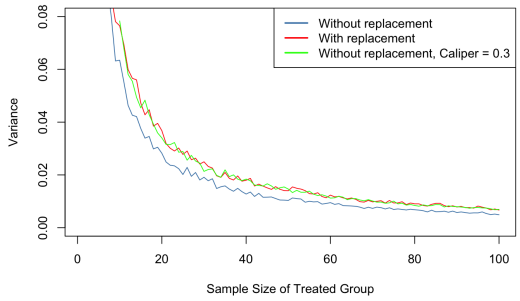
However for the far distributions, sampling without replacement gave the worst results. The caliper method performed better when the sample size was over 40 and it could be used. For smaller samples matching with replacement gave the best results although these were still biased.



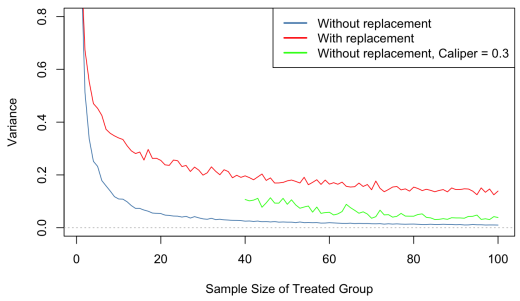
(a) *Bias for close distributions as n increases*



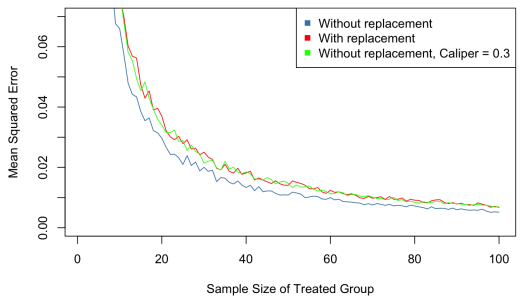
(b) *Bias for far distributions as n increases*



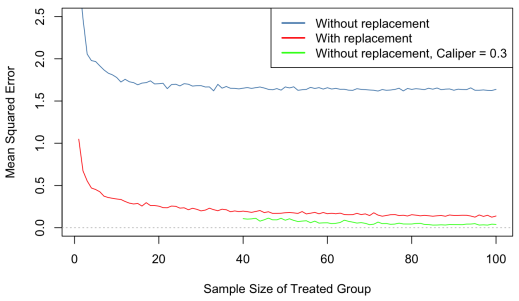
(c) *Variance as n increases for close normal distributions*



(d) *Variance as n increases for far normal distributions*



(e) *MSE as n increases for close normal distributions*

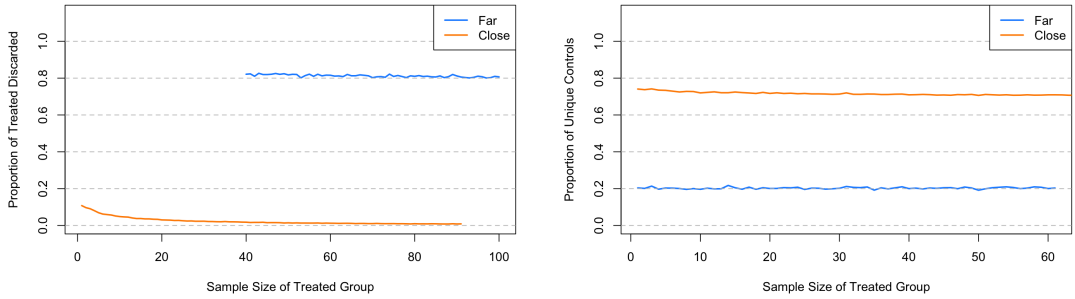


(f) *MSE as n increases for far normal distributions*

Figure 16: *Changes in Bias, Variance, and Mean Squared Error (MSE) of the estimate obtained by matching with Three Matching Methods for both high and low overlap between the treated and control groups as n increases.*

Matching Method	Bias (n = 1000)	Variance	Mean Squared Error
Without replacement	0.0034007348	0.0004238423	0.0034009145
With replacement	0.0010977321	0.0007002738	0.0010982225
Propensity score with Caliper = 0.3	0.0006781846	0.0005172175	0.0006784521

Table 4: *match quality when n=1000*



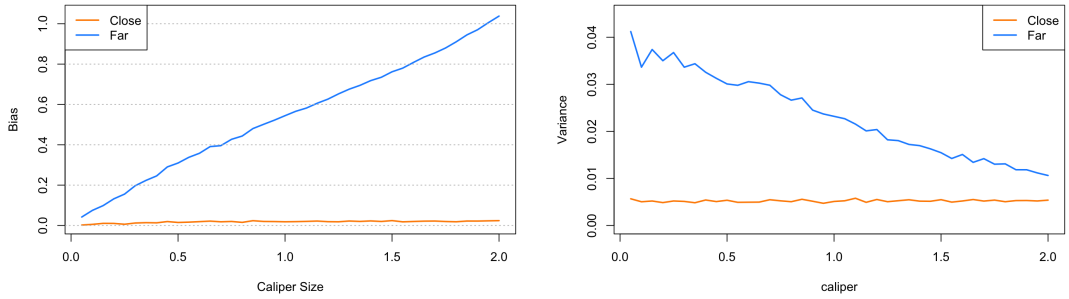
(a) *Proportion of Treated units discarded*

(b) *Proportion of unique controls*

Figure 17: *Figure (a) shows changes in the proportion of treated units discarded when matching using a caliper. Figure (b) shows changes in the proportion of unique control units when matching with replacement.*

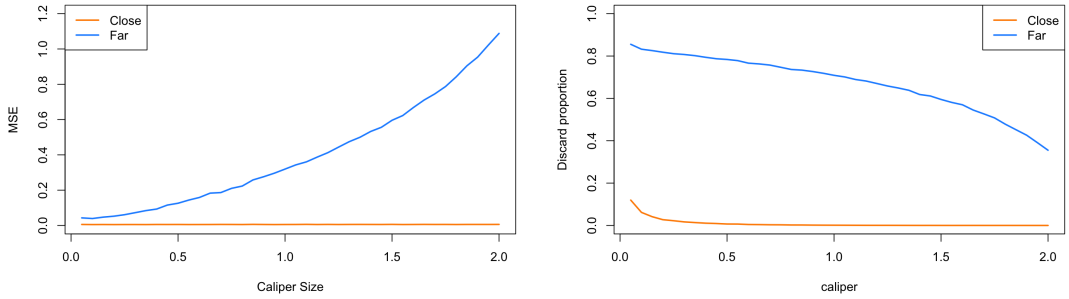
4.2 Effect of caliper size on match quality for normal data

PSM matching with a caliper was applied to data drawn from both near and far normal distributions. In figures 18a and 18b it can be seen that for close distributions with high levels of common support the caliper size has little effect on the bias or the variance. This is as expected as in the absence of interaction the overlap is the main cause of bias and of discarded treated units leading to higher variance.



(a) *Bias vs Caliper Size*

(b) *Variance vs Caliper Size*



(c) *Mean Squared Error vs Caliper Size*

(d) *Proportion of Treated Discarded vs Caliper Size*

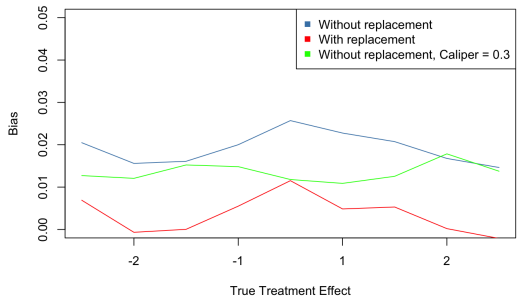
Figure 18: *Effects of changing caliper size for close and far normal distributions*

Consequently, the MSE is also affected by the caliper size for the far distributions only, with an increase in MSE as caliper size increases.

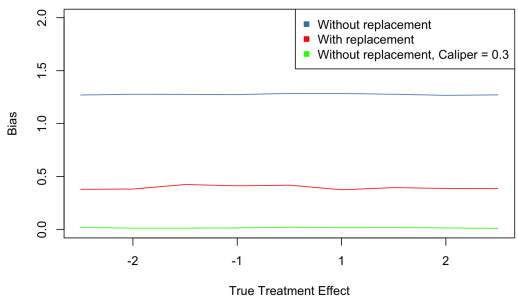
This suggests that the bias component of the MSE has more influence than the variance component, and using a very small caliper can help to mitigate the effects of low overlap. Removing points which violate the positivity assumption by using a smaller caliper appears to improve the estimates even though a very high proportion of treated units are discarded.

4.3 Effect of treatment effect size on match quality for normally distributed data

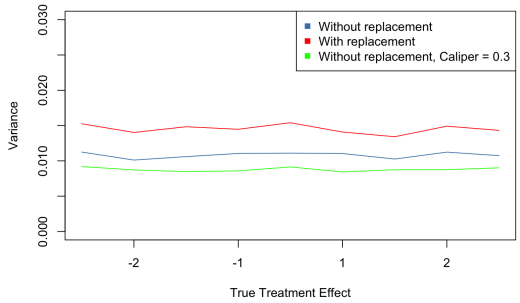
As with the uniformly distributed data, the size of the treatment effect does not affect the bias, variance or MSE of the estimate (19). This is as expected, as the true treatment effect is subtracted from the estimated effect in order to calculate the bias.



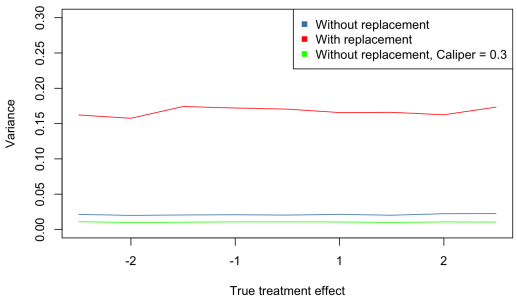
(a) *Bias vs treatment effect for close normally distributed data*



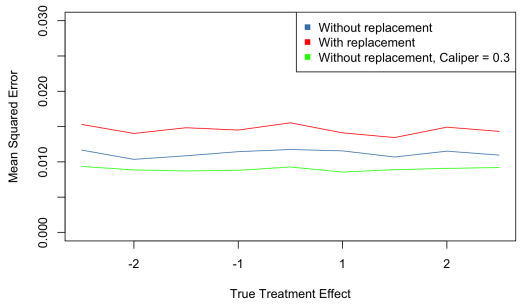
(b) *Bias vs treatment effect for far normally distributed data*



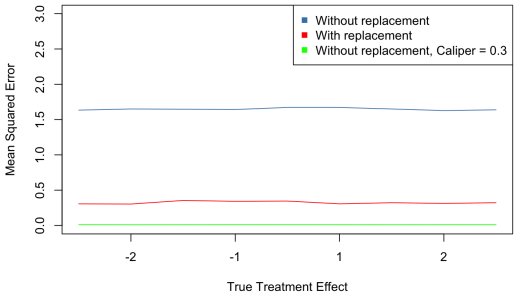
(c) *Variance vs treatment effect for close normally distributed data*



(d) *Variance vs treatment effect for far normally distributed data*



(e) *MSE vs treatment effect for close normally distributed data*



(f) *MSE vs treatment effect for far normal data*

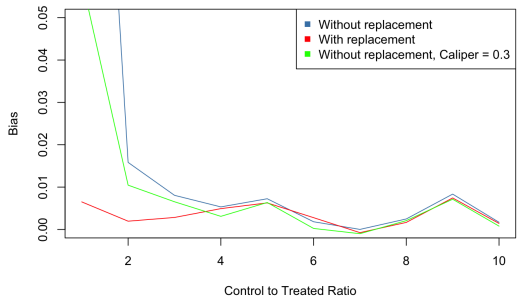
Figure 19: *Effects of true treatment effect on bias, variance and mean Squared Error of the estimated treatment effect for 3 matching methods for normal data with both close and far distributions.*

4.4 Effect on match quality of changing control to treated ratio for normally distributed data

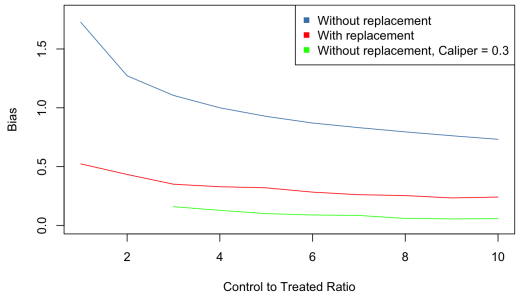
When the distributions are close, matches made using sampling with replacement give the lowest bias when the ratio of control to treated units is lower, however after the ratio is over $3:1$, there is no noticeable difference (figure 20a). For the far distributions however the matches sampled without replacement have higher levels of bias for all ratios, although they decrease slightly as the ratio increases and better matches become available (figure 20b).Matching without replacement on the far distributions gives very biased results with bias for the control to treated ratio of $2:1$ greater than the size of the targeted effect. It can

also be seen that the absolute scale of the bias is much larger for the far distributions, for all types of matching.

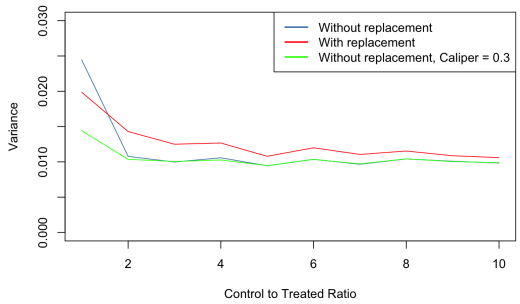
For close distributions, sampling without replacement gives the highest variance when the ratio of controls to treated units is ≥ 2 (figure 20c). For the far distributions the variance of the estimates made using matching without replacement are much higher than with with other two methods, as fewer good matches are available (figure 20d). Caliper matching was not possible for the data from far distributions when the ratio of controls to treated units was < 2. The variance of the estimates obtained from all three matches are much smaller for the close distributions than for the far distributions.



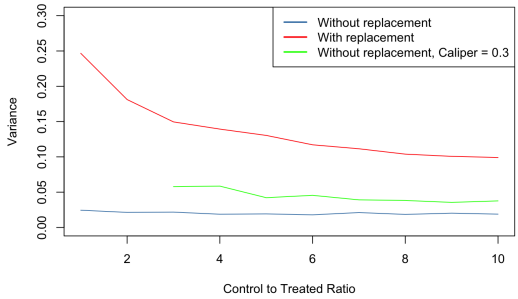
(a) *Bias vs Ratio close normal distributions*



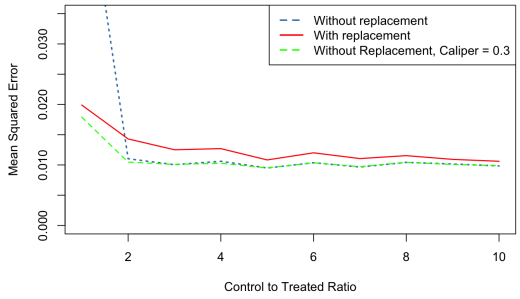
(b) *Bias vs Ratio for far normal distributions*



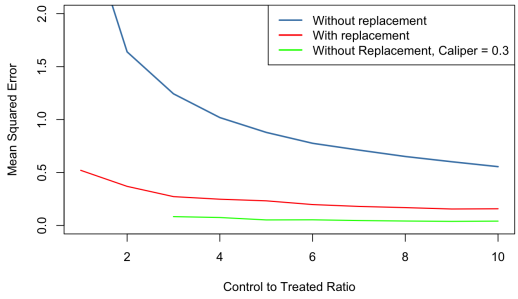
(c) *Variance vs Ratio for close normal distributions*



(d) *Variance vs Ratio for far normal distributions*



(e) *MSE vs Ratio for close normal distributions*



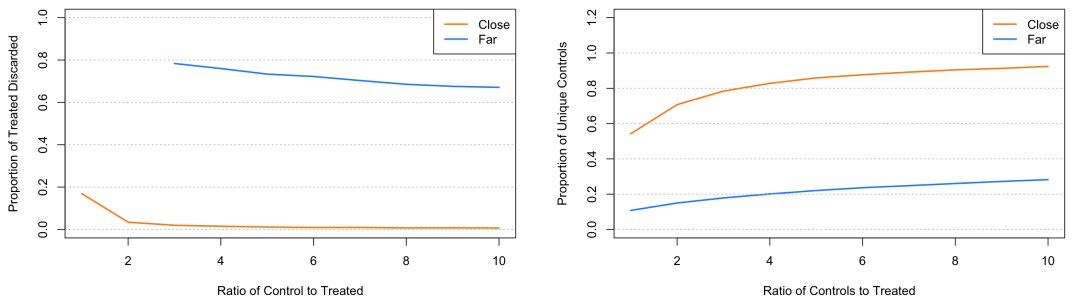
(f) *MSE vs Ratio for Far normal distributions*

Figure 20: *Bias, Variance and Mean Squared errors of the estimates obtained when using 3 types of matching methods as the ratio of controls to treated changes. Close distributions mean the treated and control groups are similar, Far distributions means the means are 3 sd apart.*

For the close distributions, as the bias was low, the variance had the greatest effect on the MSE. With a ratio of $1:1$, the MSE was highest when matching with sampling without replacement (figure 20e). For all other ratios sampling with replacement had slightly higher MSE. For normal distributions, matching using sampling without replacement and PSM matching using a caliper were equivalent for ratios of control to treated ≥ 2. For the far distributions one the other hand, the larger biases contributed more to the MSE (figure 20f). This resulted in very high MSE for matches made by sampling without replacement. Although this decreased for higher ratios, it was still much higher even when the ratio was $10:1$. Matches made with replacement performed better, and there was a slight improvement as the ratio increased.

When matching using PSM with a caliper, it can be seen in figure 21a that for the matches made using close distributions there are very few discarded for ratios ≥ 2. For the far distributions however almost 80% of treated units are discarded for a ratio of $3:1$, and for a ratio of $10:1$ there are still over 60% discarded. This suggests that the ATT cannot be estimated for the far distributions, however for the close distributions, the ATM will be close to the ATT when the ratio is over 2.

When matching with replacement, it can be seen in figure 21b that the proportion of unique controls increases with the ratio. It is also clear that many more unique controls are selected when the distributions are close together. The low proportion of unique controls used for the far distributions does not effect the estimation of the ATT as all treated units are retained. The low number does increase the variance of the estimate, and therefore the MSE.



(a) *Treated Discarded vs Ratio Changes*

(b) *Proportion of Unique Controls vs Ratio*

Figure 21: *When changing the ratio of treated to control units for close and far normal distributions, (a) shows changes in the number of treated units discarded when matching using a caliper. (b) Changes in the proportion of unique controls matched when matching with replacement.*

4.5 Effect of changing the common support proportion for normally distributed data

For overlapping normal distributions, instead of changing the overlap directly, the distance between the means was altered, giving distributions $0, 1, 2$ and 3 standard deviations apart.

When the standard deviation is 0 and the distributions are identical and the treatment effect is constant then there is no bias (figure 22a). As the distance between the means of the groups increases the bias does also. It increases most for matches sampled without replacement.

When the difference between the means is ≤ 1 standard deviation, then the variance is similarly low for all methods (22b). For greater distances it increases most for estimates from matches sampled with replacement, and slightly for estimates made with PSM and a caliper.

The overall mean squared error also increases with the overlap for all methods (figure 22c), and the increase is much higher for the estimates from matches sampled without replacement. The MSE of the estimates from matches sampled with replacement, and of the estimates from PSM with a caliper are similar when the distributions are ≤ 1 standard deviation apart. For greater distances the estimates from matches sampled with replacement are slightly higher.

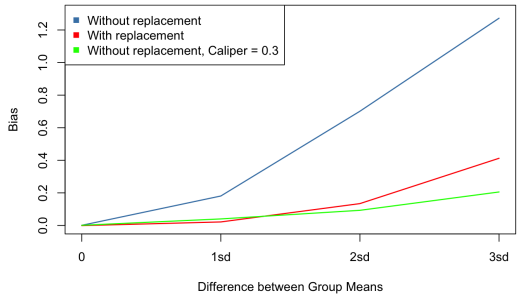
When matching with replacement the proportion of unique controls used declines as the distance between the group means increases (figure 22d). This indicates that while all treated units are matched, many are matched to the same controls. When the distance is 3 standard deviations the proportion of unique controls is less than 20% meaning that on average > 5 treated units are matched to one control. This is the cause of the higher variance when matching with replacement.

When using PSM matching and a caliper of 0.3, the proportion of treated units discarded also increases with the distance between the means. This means that as the distance between the means increases, the estimate found becomes further removed from the target estimand the ATT.

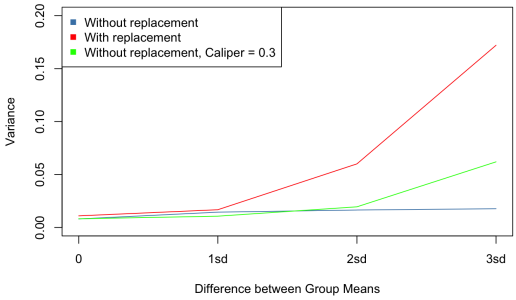
4.6 Effect of sample size on calipers

Due to the results in section 3.3, additional experiments were performed to investigate the effects of the sample size on calipers. PSM was performed on uniformly distributed data for sample sizes from 10 to 1010, with a control to treated ration of $1:1$ and increments of 50. The logit scale caliper of 0.3 was transformed to the scale of the original covariate by multiplying it by it's standard deviation.

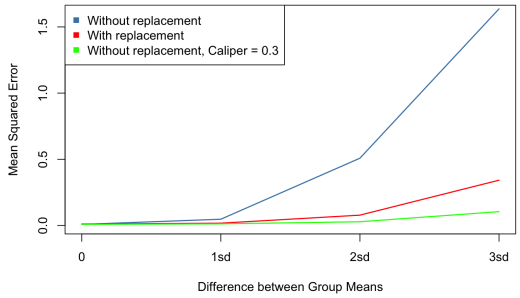
It was found that for smaller sample sizes $\leq 100)$ the effectual caliper on the original covariate scale was slightly larger.



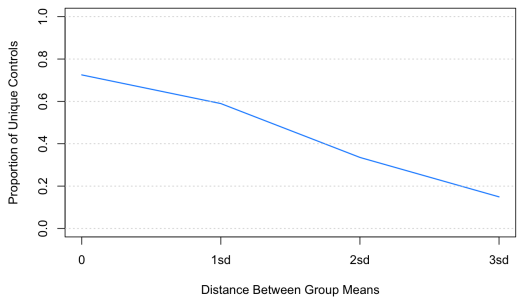
(a) *Bias as overlap changes*



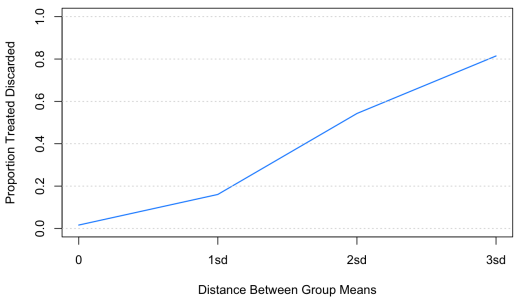
(b) *Variance as overlap changes*



(c) *Mean squared error as overlap changes*



(d) *Proportion of unique controls (with replacement)*



(e) *Proportion of treated discarded (PSM with caliper)*

Figure 22: *Effects of changing the overlap of normally distributed control and treated groups. The standard deviation of the normal distributions is* $\sqrt{\frac{1}{12}}$ *for both distributions.*

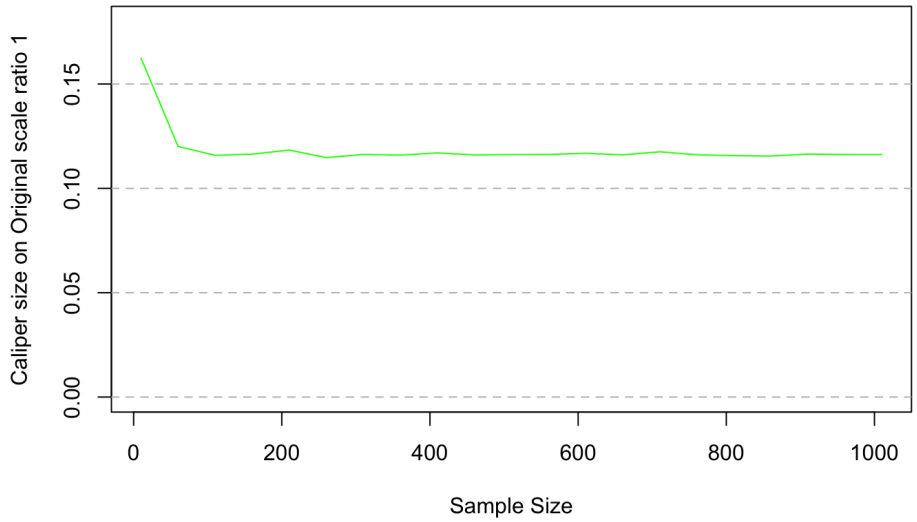


Figure 23: *Caliper of* 0.3 *on the scale of the original covariate as sample size increases when using PSM matching and the ratio of control to treated units* $= 1:1$.

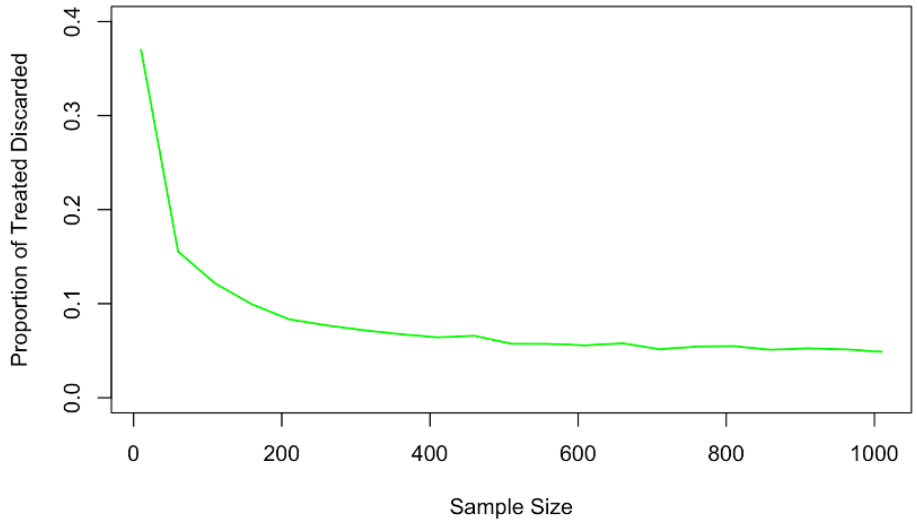


Figure 24: *Proportion of treated units discarded as sample size increases when using PSM matching and the ratio of control to treated units* $= 1:1$.

5 Discussion and recommendations

This study considered the effects of many parameters on the quality of estimates obtained using three different matching methods. Consequently there are many separate conclusions which can be combined to give an overview of the results.

- In using different levels of overlap with the uniformly distributed data, the assumption of positivity was deliberately broken, as for certain covariate values, it was not possible for both control and treated units to occur. As was seen in (figure 9), this leads to bias, especially when matches are made using nearest neighbour matching without replacement. In the experiments using normally distributed data, the assumption of positivity holds theoretically as there is no covariate value which cannot contain both treated and control units. In reality, the probability of both treated and control units occurring is practically 0 for some covariate values. This is especially true for smaller sample sizes. This situation becomes more likely as the means of the normally distributed control and treated groups become further apart. The result is that when the sample size is small, or when the group means are highly separated, the positivity assumption is not truly met and the estimates can be biased.

- There are two main sources of bias, bias from the lack of overlap and bias due to overlap combining with interaction effects. These biases can be different, cancelling each other out, or can be additive. In the interaction scenario used here negative bias was introduced from the interaction effect, and positive bias from the overlap. This caused the biases to cancel for the matches sampled with and without replacement. PSM with a caliper however, removed most of the positive bias caused by the overlap effect, leaving only the negative bias of the interaction. In other scenarios this could easily go another way. The overlap effects bias, and when interaction is present it can have an even stronger effect and in more unpredictable ways. As interaction effects are common in real data, the degree of the overlap becomes more important.

- In many situations, propensity score matching using a caliper gives the "best" result, producing the lowest mean squared error on the simulated data. However Caliper matching can fail when the distributions of treated and control groups are highly separated, especially for low sample sizes. For normally distributed data with far apart means and therefore little overlap, if $n = 50$ then the algorithm does not converge and no matches are found for any size caliper. This is due to the logit scale of the propensity scores used to calculate the caliper. When the probabilities of an individual being in the treatment or control group approach 1 or 0, on the logit on which the calipers is applied they approach plus and minus infinity, so no caliper is ever big enough to match them.

- In situations where caliper matching causes treated units to be discarded, then it should be noted that the estimate produced is no longer the desired estimand the ATT. It is instead the estimate of what can be called the Average effect of treatment on the matched or ATM. When treated units with certain covariate values are left unmatched, it is no longer possible to make inferences about the effect of the treatment on individuals with covariates in that range. In this case, it cannot be said that the treatment causes X effect, but rather that the treatment causes X effect for individuals with covariate values between a and b.
- Propensity score matching with a caliper on far normal distributions discards a high proportion of treated units. This occurs even when there is a high ratio of controls to treated. (figure 21a). A higher proportion of treated units discarded means that the estimate is further from the targeted ATT. Although it initially appears (e.g. in figure 20b) that PSM matching with a caliper gives the best estimate, there are two caveats. Firstly, as can be seen from the high proportion of discards, the estimate can no longer be called the ATT. Secondly, in this simulation there was a constant treatment effect. This meant that the correct estimate of the ATT could be obtained matching on only a small proportion of the covariate values. This would not apply in the common situation where there is an interaction effect.
- As noted in section 4.6 the effectual caliper on the original covariate scale was slightly larger when n is smaller. This indicates that the actual size of the caliper is sample size dependent. This could be due to the standard error of the logit being smaller with smaller sample sizes as points might not occur at the extreme ends of the distribution. A larger caliper should result in higher bias, however it is balanced by the higher proportion of treated units discarded for smaller sample sizes. In this simulation where there is a constant effect and no interaction the high number of discards does not affect the estimate, this produces the strange situation of lower bias for smaller samples when the ratio is $1:1$.
- The propensity scores were estimated in R using a generalized linear model with a binomial family and a logit link function. As logistic regression is commonly used, this could leave propensity score estimates vulnerable in situations where logistic regression fails. One of these situations is when there is perfect separation of the groups as it is unclear where to put the decision boundary. This situation can in these simulations when the uniform distributions are used and the overlap and sample size are small. It also happens to a lesser extent in the normal simulations when the distributions are far apart. Smaller sample sizes make an unclear decision boundary more likely, especially when the distributions are highly separated. Smaller sample sizes also cause the odds ratios and therefore the logit to be biased away from the null (Nemes, Jonasson, Genell, & Steineck,

2009). This can explain the strange results with the calipers when ratio $= 1$ and n is small.

- It should be noted that one of the reasons for the popularity of PSM is in simplifying complex multidimensional datasets down to a single number. This makes the matching process much faster and easier to implement. A limitation of this study was that only a single covariate was used so this did not apply. However, matching on many covariates may not be a good idea in any case, as less important covariates may obscure more important ones. Using PSM may make researchers inclined to throw in as many as possible without consideration of their usefulness, but this could lead to poorly calculated propensity scores and sub-optimal matches.
- In the literature it has been recommended that a caliper size of 0.2 is the best (Austin, 2011). The results in these simulations indicate that the ideal caliper size, especially for smaller sample sizes and low levels of overlap, is dependent on the bias variance trade off, and as such can be larger than 0.2 (figure 6c). In this simulation the sample size of the treated was 50 with a control to treated ratio of $2:1$, and for low overlap levels the best MSE was obtained with calipers between 0.3 and 0.5, while for higher levels of overlap the caliper size had very little effect on the quality of the estimates. These results indicate that the best caliper size varies according to sample size and the level of overlap, and that one size does not fit all datasets.

Guidance for matching

Positivity as represented by overlap in this study is very important and should be checked. The proportion of treated units discarded is very important when using propensity score matching, this should be checked and reported as it indicates how far the estimate may be from the target estimand. Likewise, if matching with replacement, the number of unique controls used should also be checked, especially when there is low sample size, overlap or ratio of treated to control units. If using a caliper it is also important to be aware that a caliper of a fixed size on the logit scale is not always the same on the original scale. Additionally, the sample size should be taken into consideration when selecting the caliper as for smaller samples calipers greater than the common 0.2 produce better estimates.

The goal of matching is to improve covariate balance and allow better estimates to be made. When there is high overlap of the treated group and positivity exists everywhere then this can work well. When there is not, then bias is introduced. When there is low overlap and interaction then even more bias is introduced. When the overlap is small, valid conclusions can only be drawn for covariate values where positivity exists and should not be extrapolated beyond that.

Ultimately if the data set is too small or if the positivity assumption is stretched too far then no matching technique is going to help and it is necessary to gather more data or attempt a different modelling method.

Conversely, when there is a large data set with high overlap between the groups then all methods will give reasonable results. In this situation matching with replacement should allow the lowest bias while still providing the target estimand.

Possible future research

This study used a single covariate, future studies could consider performance of multiple covariates. The study could also be extended to include other matching algorithms, for example coarsened exact matching (CEM).

The effect on bias of the empirical overlap in real data sets could also be examined. This could be done either by using a large dataset and cross validation or by making a realistic simulation based on real data.

In this study, some of the issues relating to propensity score matching seemed related to the use of logistic regression. In the future other methods of estimating the propensity score such as tree based methods could be investigated.

6 Conclusion

There is no one best matching method for every situation. Propensity score matching may seem at first to be a silver bullet, simplifying multivariate datasets and providing accurate estimates with low bias. However as overlap decreases and the positivity assumption is stretched to breaking the number of treated units discarded take the estimate ever further from the target estimand. PSM also performs badly when both the sample size and the control to treated ratio are very small.

Matching with replacement solves the issues of poor overlap and low ratio leading to few good matches by reusing control units. This is effective and produces reasonable results when the sample size is sufficiently large. For small sample sizes and low overlap the limited number of unique control values means that the variance is high and this is not compensated for sufficiently by the reduction in bias. The number of unique controls used is directly related to both the ratio of treated to control units and to the overlap.

Matching without replacement means that the estimate is very vulnerable to bias caused by poor overlap, and this is especially true for smaller control to treated ratios.

7 R version and link to code files

The simulations and graphs for this thesis were created in R studio version 1.4.1106

The files are available to view on GitHub.

References

Ali, M. S., Groenwold, R. H., Pestman, W. R., Belitser, S. V., Roes, K. C., Hoes, A. W., ... Klungel, O. H. (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiology and drug safety*, *23*(8), 802–811.

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.

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Greifer, N. (2022). *Vignette 'assessing balance'.* cran.r-project.org.

Hernán, M. A., & Robins, J. M. (2018). *Causal inference.*

Ho, D., 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.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. doi: 10.18637/jss.v042.i08

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, *20*(1), 1–24.

King, G., Lucas, C., & Nielsen, R. A. (2017). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science*, *61*(2), 473–489.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, *27*(4), 435–454.

Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology*, *9*(1), 1–5.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software, Forthcoming*.

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.

Westreich, D., & Cole, S. R. (2010, 02). Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, *171*(6), 674-677. Retrieved from https://doi.org/10.1093/aje/kwp436 doi: 10.1093/aje/kwp436

List of Figures

List of Tables