Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2021;7(3): 251-262

Original Article

Comparison of Nearest Neighbor and Caliper Algorithms in Outcome Propensity Score Matching to Study the Relationship between Type 2 Diabetes and Coronary Artery Disease

Sara Sabbaghian Tousi[1], Hamed Tabesh[2], Azadeh Saki[1,3*], Ali Tagipour[3,4], Mohammad Tajfard[3,5]

[1]Department of Epidemiology and Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.
[2]Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[3]Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.
[4]Department of Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.
[5]Department of Health Education and Health Promotion, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran.

ARTICLE INFO

Received 15.04.2021
Revised 19.07.2021
Accepted 22.08.2021
Published 25.09.2021

Key words:
Propensity score matching;
Caliper algorithm;
Nearest neighbor algorithm;
Diabetes;
Coronary artery disease

ABSTRACT

Introduction: Propensity score matching (PSM) is a method to reduce the impact of essential and confounders. When the number of confounders is high, there may be a problem of matching, in which, finding matched pairs for the case group is difficult, or impossible. The propensity score (PS) minimizes the effect of the confounders, and it is reduced to one dimension. There are various algorithms in the field of PSM. This study aimed to compared the nearest neighbor and caliper algorithms.

Methods: Data obtained in this study were from patients undergoing angiography at Ghaem Hospital in Mashhad, between 2011-12. The study was a retrospective case-control using PSM. In total, 604 patients were included in the case and control groups. A logistic regression model was used to calculate the propensity score and adjust the variables, such as age, gender, Body Mass Index (BMI), systolic blood pressure, smoking status, and triglyceride. Then, the Odds Ratios (ORs) with 95% Confidence Intervals (CIs) for the raw data and two matching algorithms were determined to examine the relationship between type 2 diabetes and coronary artery disease (CAD).

Results: Propensity score in the nearest neighbor and caliper algorithms matched the total number of 604 samples, 200 and 178 pairs, respectively. All variables were significantly different between the two groups before matching ($P<0.05$). The gender was significantly different between the two groups after matching using the nearest neighbor algorithm ($P=0.002$). No variables created a significant difference between the two groups after matching with the caliper algorithm.

Conclusion: Bias reduction in the caliper algorithm was greater than for the nearest neighbor algorithm for all variables except the triglyceride variable.

Introduction

The rates of diabetes are increasing all over the world. The scientists estimate that the number of diabetics will go up dramatically in the next years and will reach the number of 592 million

*.Corresponding Author: sakia@mums.ac.ir

by 2035. The main system that affected by diabetes is cardiovascular and causes death in more cases. Patients suffering from diabetes are disposed to more severe cardiovascular diseases and have greater complication rates than non-diabetic patients. The most common cardiovascular diseases are coronary artery disease (CAD). There is a strong relationship between CAD and type 2 diabetes. So, diabetes is considered a CAD risk equivalent. This means that diabetic patients are at risk of having coronary disease similar to non-diabetic patients, who had one before. Many factors contribute to the appearance of CAD in diabetes type 2 patients and only 25% of these are already known.[1]

Due to the fact that there are several confounding variables in the relationship between type 2 diabetes and CAD. We need to use the propensity score matching method because if we try to match these confusing variables, we may encounter over-matching that we may not find any matching control for a subject in the case group. Therefore, to find controls with the same score, we must use the propensity score matching method.

Matched case-control studies are one of the most common approaches for studies in health sciences. One of the most critical points in designing these studies is the distribution of similar confounders in the case and control groups.[2] When the number of confounders is high, there is often a problem of matching, and it is difficult or impossible finding the matched pairs for the case group.[3] In case-control studies, matching is done on the response variable (unlike cohort studies where matching is done on exposure). Therefore, in calculating the propensity score, the logistic model of the relationship between confounding variables and response variable should be used.

Instead, The propensity score is calculated using the logistic regression model based on the outcome (patient=1 and healthy=0) conditional on the observed confounders.[4] Indeed, PSM is used in case-control studies to evaluate the results. PSM is the best way to overcome selection bias and confounding factors in observational studies, by creating a balance between the two groups.[5] There is a wide range of different methods for forming matched pairs, such as optimal matching, nearest neighbor matching, and caliper matching.[6] The nearest neighbor algorithm is the most common method used for this in medical science, although the caliper width is not constant in this algorithm.[7, 8] In addition, previous studies have investigated performance calipers with different widths.[9] While there are few studies that have compared the performance of different algorithms of the propensity score matching.[6] But now these studies are on the rise. The aim of this study is to compare the performance of two propensity score matching algorithms called the nearest neighbor matching and the caliper matching to find an appropriate matched control group to investigate the relationship between diabetes and coronary artery disease with considering the confounding variables in this relationship.

Materials and Methods

We used the method of PSM in a case-control study to investigate the relationship between type 2 diabetes and coronary artery disease in candidates for angiography in Ghaem hospital in Mashhad from September 2011 to August 2012. A total of 604 patients were included in the case and control groups. The case

group consisted of 200 patients who had been diagnosed with 2 to 3 coronary artery stenosis by angiography. The control group consisted of 404 healthy people, over 18 years, who were referred for regular medical examinations. After obtaining informed consent demographic information was recorded, that included: age, gender, smoking status, history of diabetes (or not). The following parameters were measured: serum triglycerides, systolic blood pressure, body mass index, as previously reported.[10, 11] We used PSM to select the appropriate matched pairs between the two groups with, or without coronary artery disease.

First, we checked confounder variable entry conditions and evaluated the balance between the case and control groups. We used statistical methods such as the chi-squared (χ^2) test for variables. If at least one variable creates an imbalance between the two groups, the chi-squared test will be statistically significant,[12] standard difference estimate for a comparison of the means or medians of continuous covariates and the distribution of their categorical counterparts between case and control subjects. For a continuous covariate, the standardized difference is defined is defined as:

$$d = \frac{\left|\bar{x}_{case} - \bar{x}_{control}\right|}{\sqrt{\frac{s^2_{case} + s^2_{control}}{2}}} \times 100$$

$\bar{x}_{case}$ and $\bar{x}_{control}$ are the sample mean of confounder variables covariates in the case and control groups, respectively. s^2_{case} and $s^2_{control}$ are the sample variance of covariates in the two groups, respectively. For dichotomous confounder, the standardized difference is defined as:

$$d = \frac{(\hat{p}_{case} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{case}(1-\hat{p}_{case}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}}$$

$\hat{p}_{case}$ and $\hat{p}_{control}$ are the prevalence or mean of the dichotomous variable in case and control groups, respectively. Although there is no universal agreement criterion as to what threshold of the standardized difference can be used to represent an important imbalance, a standard difference that is less than 0.1 has been taken to represent an insignificant difference in the mean or prevalence of a covariate between groups.[13]

The propensity score is estimated using a logistic regression model. The numerical value of this score is between zero and one, and it is defined as a probability of being in the case or control group on the basis of confounder variables: $e(X_i) = \Pr(Z_i = 1 \mid X_i = x_i)$

Were Z_i is binary outcome variable (1=disease/0=non-disease).

$e(X_i)$ is an estimation of the propensity score for the i^{th} individual, where X_i represents the vector of confounder variables for the i = individual, which will be matched in two groups.[4] In order to achieve normality, the logit of the propensity scores are usually used instead of the propensity scores $\hat{e}(X_i)$, where β is a vector of the regression coefficients [14, 15]. $\ln(\frac{e(X_i)}{1-e(X_i)}) = \beta X_i$

The matching has the best performance when there is an appropriate overlap between the logit of estimation of the propensity scores between the case and control groups. Therefore, we examined the overlap between the two groups using a box plot.[16]

We then implemented the nearest neighbor and caliper algorithms with a 1:1 ratio (an individual matched from the control group with an individual from case group) and without replacement.

The nearest neighbor algorithm is a method

that matches the two groups based on the nearest distance. If the absolute distance between their propensity scores is the smallest value, the jth person with the propensity score is in the control group a proper match for the ith individual with the propensity score in the case group.[17] If multiple subjects in the control group have equally close propensity scores to the propensity score of the sample subject in the case group, one of those is selected at random.[18]

$$d(i,j) = \min_j \left\{ \left| e(X_i) - e(X_j) \right| \right\}$$

The caliper algorithm uses the absolute distance of the propensity scores of individuals in groups less than a specified caliper.[6]

$$d(i,j) = \min_j \left\{ \left| e(X_i) - e(X_j) \right| < \varepsilon \right\}$$

$e(X_i)$ and $e(X_j)$ are the propensity scores of the individual in the case and control group, respectively. $\varepsilon < 0.25\,\sigma_p$ is a pre specified caliper.[19] σ_p is the standard deviation of the logit of estimation of the propensity scores, where σ_i^2 is the variance of logit of the propensity score in the i[th] group. Range a is allowed to change from 0.05 to 2.50 in increments of 0.05.

$$\sigma_p = a\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$$

If the variance of the logit of the propensity score is similar in both groups, the caliper with a width of 0.2 removes approximately 99% of bias due to measured confounder variables.[9]
We evaluated the mean difference for all variables in the case and control groups before and after the matching with the two algorithms as well as the bias reduction percentage for each variable after the matching. We expected that the mean difference and bias would be reduced after matching with the algorithms in comparison with before matching, which shows the matching algorithms are useful.
The main purpose of matching is to reduce selection bias by increasing the balance between the case and control groups. For a continuous covariate, the bias reduction percentage is defined as:

$$\frac{(\bar{x}_{Acase} - \bar{x}_{Acontrol}) - (\bar{x}_{Icase} - \bar{x}_{Icontrol})}{\bar{x}_{Icase} - \bar{x}_{Icontrol}} \times 100$$

where $\bar{x}_{Icase}, \bar{x}_{Icontrol}, \bar{x}_{Acase}$ and $\bar{x}_{Acontrol}$ are the mean of confounder variables in the case and control groups before and after matching, respectively. For dichotomous confounder, the standardized difference is defined as:

$$\frac{(\hat{p}_{Acase} - \hat{p}_{Acontrol}) - (\hat{p}_{Icase} - \hat{p}_{Icontrol})}{\hat{p}_{Icase} - \hat{p}_{Icontrol}} \times 100$$

where $\hat{p}_{Icase}, \hat{p}_{Icontrol}, \hat{p}_{Acase}$ and $\hat{p}_{Acontrol}$ are the prevalence or mean of the dichotomous variable in the case and control groups before and after matching, respectively.[20]
A fixed value of the bias reduction is unclear, but 80% bias reduction may be a reasonable and sufficient.[21] This value, or greater indicates that many appropriate samples have been matched.[22]
We evaluated PSM algorithms quality using statistical tests and graphical methods. In this study, we used paired t-test for quantitative variables and McNemar test for qualitative variables. Hypothesis H_0 (there is no significant difference between the variables of the two groups) were tested against the hypothesis H_a (there is a significant difference between the variables of the two groups).[8]
One of the graphical methods for checking the quality of algorithms is the distribution histogram of the propensity scores before and after matching. The improvement in the

distribution of propensity scores is evaluated between the case and control groups by comparing these two histogram plots.[23] Important parameters to determine the fit are not only the shape but also is the degree of overlap between the two distributions that known as the common support region (by examining the Y axis).[16] Matching is best when there is a common support region. The Quantile-Quantile (QQ) plot compares the probability distributions of a confounder in the case and control group by plotting their quantiles against each other.[13]

Finally, we examined the sensitivity of the results obtained from the study for the hidden biases (for which the researcher did not recognize the confounder variables and did not enter the study).[24] Rosenbaum method is used for sensitivity analysis. If the outcome and exposure are nominal variables, the McNamar test is used in this method.[25] We performed sensitivity analysis the intervals and p-values are obtained using this test. Where u is an unobservable confounder variable, so we tested the hypothesis H_0 (u included in the study, the study is sensitive) against the hypothesis Ha (u do not include in the study, the study is not sensitive).

$$\begin{cases} H_0 : E(u) \neq 0 \\ H_a : E(u) = 0 \end{cases}$$

The existence of the Unobservable confounder variable u is investigated using the gamma index (Γ). Gamma is the sensitivity parameter and a measure to determine the robustness of results relative to the hidden bias. For each, $\Gamma \geq 1$ the boundaries are provided for the significant levels of the null hypothesis. Different gamma values are investigated for upper and lower bounds and their significant levels. Finally, a study is sensitive to hidden bias if the gamma value is close to one for change at significant levels.[26]

All statistical analyses were performed using R version 3.3.4. The level of significance was set at $P<0.05$.

Results

Using standard deviation and chi-squared test, age variable (53%) and gender and smoking status variables ($P <0.001$) created an imbalance between the two groups. Other variables (BMI, triglycerides, and systolic blood pressure) have been used as confounder variable by previous studies. As shown in Figure 1, there was an appropriate overlap in the estimated propensity scores of the two groups.

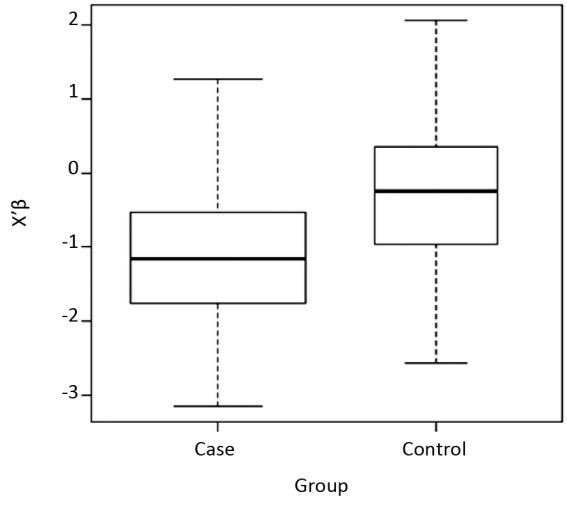

Figure 1. Overlap on fitted scores

We calculated the propensity score using a logistic regression model. CAD was fitted as the response variable versus variables such as age, gender, BMI, smoking status, triglyceride, and systolic blood pressure.

As shown in Table 1, the mean difference in the

Table1. Performance Evaluation of Matching Algorithms

variable	Before matching		After nearest neighbor matching			After caliper matching		
	mean difference	P-value	mean difference	P-value	Reduction bias (%)	mean difference	P-value	Reduction bias (%)
Age (years)	-0.29	<0.001	-0.06	0.141	77.63	-0.02	0.795	92.26
Sex (male %)	5.87	<0.001	1.34	<0.001	77.11	0.92	0.984	84.32
BMI (kg/m^2)	5.73	<0.001	1.59	0.63	72.27	-0.90	0.263	84.23
Smoking status (Non-smoker %)	15.82	0.035	8.23	0.248	47.98	5.89	0.992	62.71
Systolic blood pressure (Mm Hg)	0.25	0.022	0.23	0.525	9.17	0.35	0.909	39.91
Triglyceride (Mg/dl)	0.06	<0.001	0.01	0.225	78.26	-0.02	0.540	67.43

caliper algorithm was lower for all variables except systolic blood pressure and triglyceride variables compared with the nearest neighbor algorithm. The bias reduction percentage was increased with the caliper algorithm for all variables except the triglyceride variable in comparison with the nearest neighbor algorithm.

In addition, Table 1 shows that all variables were significantly different between the two groups before matching (P<0.05) but after applying nearest neighbor matching only the gender variable created a significant difference between the two groups (P=0.002). All the variables created the balance between the two groups after the caliper matching.

Then, the nearest neighbor and caliper algorithms 200 and 178 pairs were matched in the case and control groups, respectively.

We used a graphical approach called back-to-back histogram to assess the distributional similarity between score distributions. As shown in figure 2 (A) the propensity scores were predicted to data before matching. According to the Y-axis, there was a common support region between the case and control groups. As can be observed in this figure 2 (B) and (C), there is improvement in their common support region and also in the match between the two distributions of propensity scores after the matching with the caliper and the nearest neighbor algorithms with compared to figure 2 (A), which shows the histograms for the same data before the match. Eventually, in Figure 2, the caliper algorithm had a remarkable improvement for distribution histograms of the propensity scores in the case and control groups before and after matching with compared to the nearest neighbor algorithm. Thus, this algorithm has reduced the selection bias significantly.

According to Figure 3, after matching with two algorithms, the points were not exactly on the 45 degree line (y=x), but there was a continuous empirical distribution between the two groups. Deviations from the 45 degree line in the caliper algorithm was less than the nearest neighbor algorithm.

Gamma was 3.4 and 2.8 in the nearest neighbor and caliper algorithms, respectively. This study was not sensitive because gamma is not very close to one. Therefore, the results of

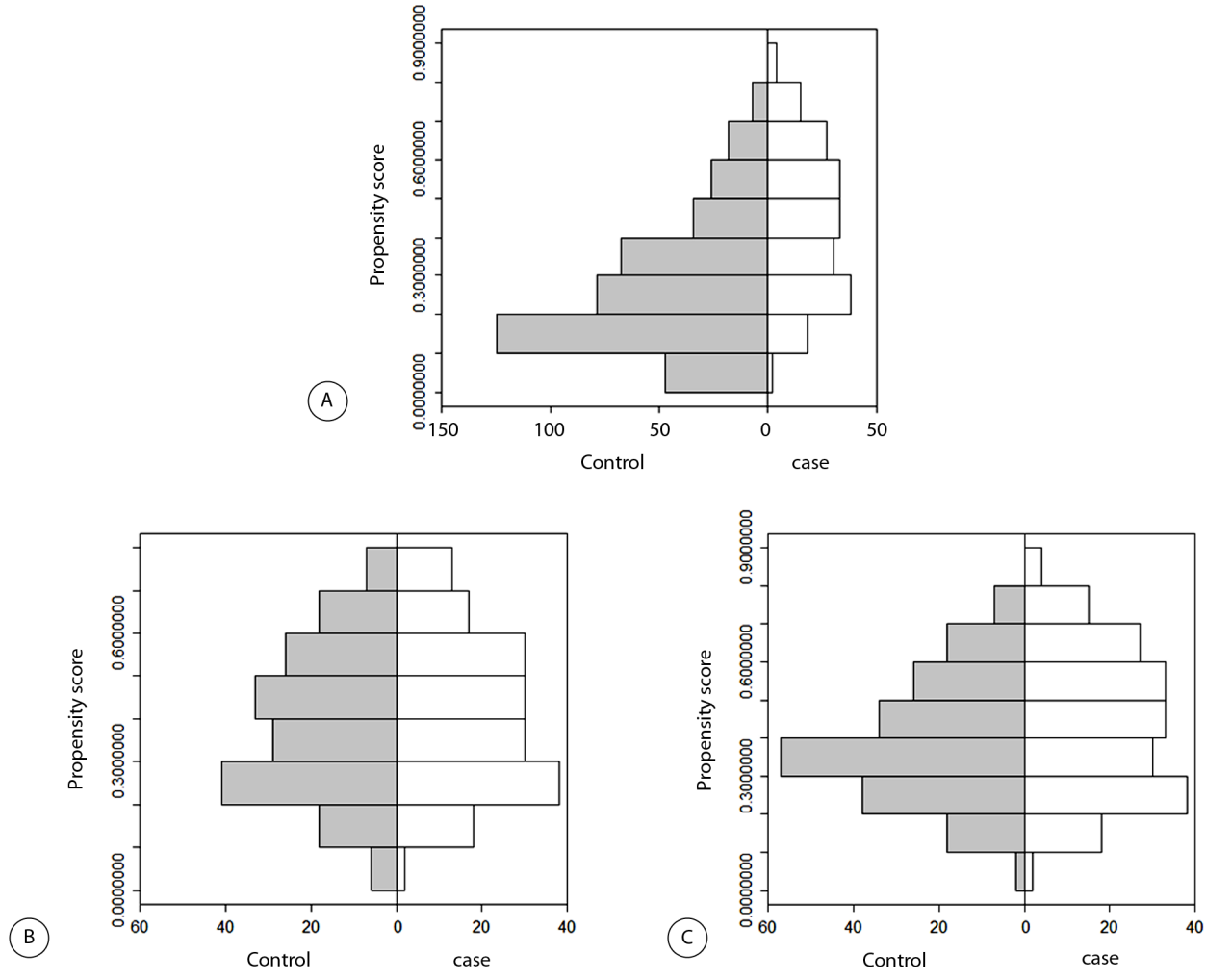

Figure 2. Histogram Distribution of the propensity scores before matching (above) the nearest neighbor algorithm (bottom right) the caliper algorithm (bottom left)

this study were robust relative to the observed confounders, and the existence of another confounder did not change the results.

After evaluating the quality of the two algorithms, the balance was shown to have a remarkable improvement for all confounder variables that create a significant difference between the two groups. The caliper algorithm performed better than the nearest neighbor algorithm in selecting appropriate matched pairs. Selection bias reduced and the matching was effective.

In Table 2, we calculated the odds ratio (OR) for raw data without adjustment and adjusted by logistic regression model and for matched data with two algorithms. It was determined that CAD was positively related with diabetes mellitus. The confidence interval of OR after matching with the two algorithms was wider than the before matching because data was in the form of the pair after matching and the sample size reduced. To determine the accuracy of real OR values on raw data, Mantel-Hansel was used and it was observed that after adjusting on the variables of age, gender, smoking status, body mass index,

systolic blood pressure, and triglyceride that were categorized, the value of OR decreases and becomes close to the value of OR obtained from the matching with the caliper algorithm. So that the odds of coronary artery disease in diabetics was 4.85 times higher than in non-diabetics in the caliper algorithm.

Discussion

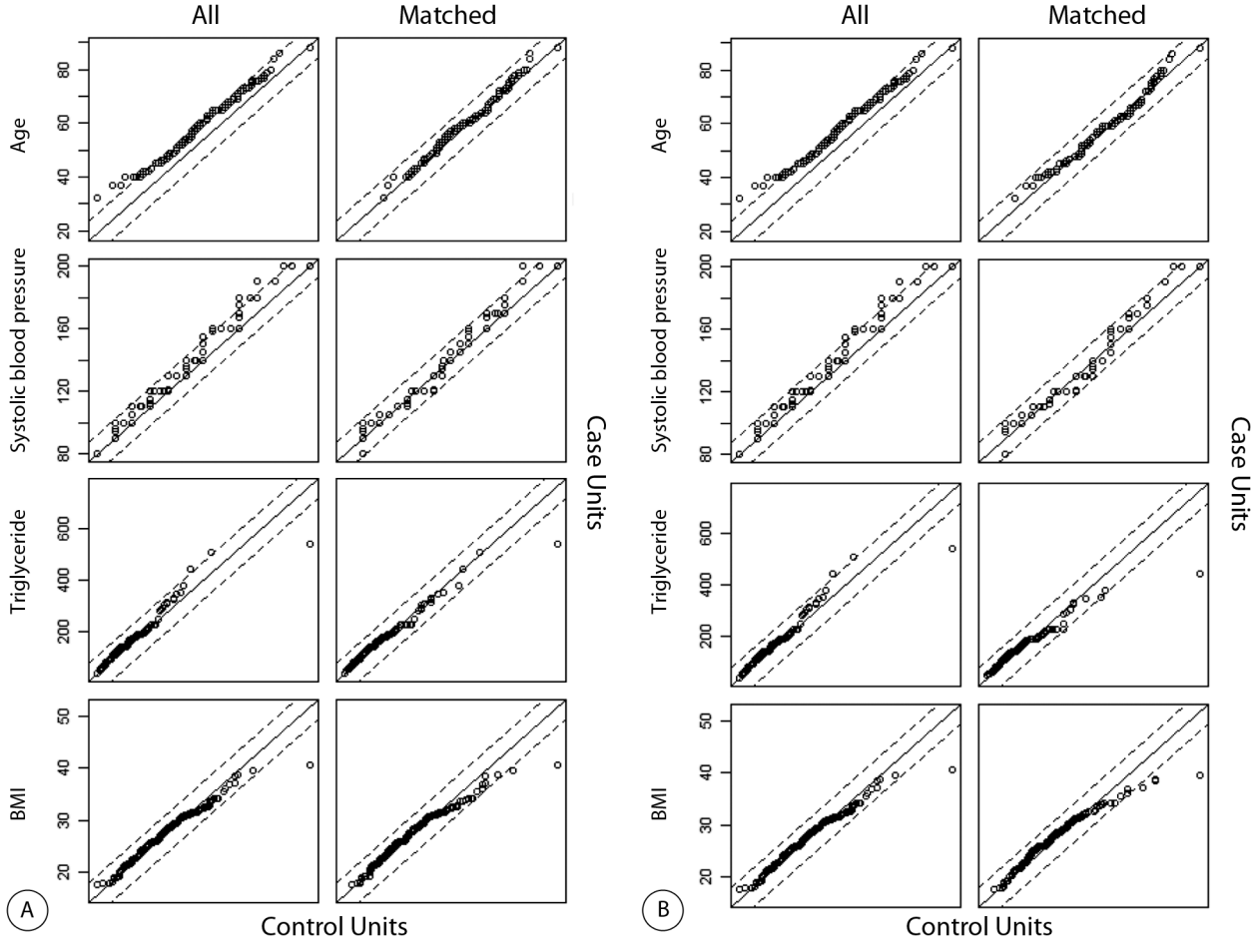

Figure 3. QQ plot of the caliper algorithm (right) QQ plot of the nearest neighbor algorithm (left)

Table 2. Comparing odds ratio before and after matching with the caliper and the nearest neighbor algorithm

	OR	OR (95%CI)
Raw data without adjustment on confounders (n = 604)	5.64	(3.89-8.16)
Raw data adjusted by logistic regression model (n = 604)	5.12	(3.37-8.78)
Nearest neighbor matching	5.6	(3.02-10.24)
Caliper matching	4.85	(2.71-8.53)

The purpose of this study was to compare the propensity scores using two nearest neighbor and caliper algorithms. Propensity score in the nearest neighbor and caliper algorithms matched the total number of 604 samples, 200, and 178 pairs, respectively. In order to perform individual matching based on study confounders, we needed a large population to select the control group. In this study, it was not possible to match any pairs by individual matching.

There must be an appropriate overlap between the distributions of the propensity scores of the two groups. If the overlap is small, there may not be sufficient numbers of individuals in the control group to be matched with all individuals group; in the case group. Therefore, PSM would not be better than matching using standard approaches. Therefore, the minimum sample size is considered for the control group 3-4 times larger than the case group.[27] In this study, the sample size was for the control group more than two times the case group.

Matching can be done using replacement, and a ratio of many to one (M:1) or one to many (1:M), but in the medical field, it is rarely used because It will be difficult to check the balance between the two groups.[28] In this study, the matching algorithms were performed without replacement and with a ratio (1: 1).

Austin performed a study of PSM, comparing different algorithms, and in which the caliper algorithm created the most balance between the two groups.[6] In this study, the selection of appropriate matching algorithm had a remarkable effect on the quality of the matched pairs, and the caliper algorithm reduced 40%-92 % of the selection bias for confounder variables.

Austin also performed a study to compare two widths of the caliper algorithm with 0.2 and 0.6 the standard deviation of the logit of the propensity score that were eliminated 99% and 90% of the bias of the variables between the two groups, respectively. Therefore, the caliper algorithm had the best performance with a width of 0.2 standard deviation of the logit of the propensity score.[29] This study presented, the use of an optimal caliper is vital for achieving appropriate matches pairs.

Sensitivity analysis is necessary to determine the robustness of the results. In the PSM method all the variables must be entered that cause the imbalance between the two groups, because the lack of a potential variable in estimating the propensity scores does not reduce the selection bias, and the effectiveness of this method is limited.[24] In a review study in the medical field, only one of the 27 articles had been reported sensitivity analysis.[26] In this study reported sensitivity analysis and the study was not sensitive because the results were robust relative to the observed confounders.

Pirracchio performed a study, and the odds ratio for a matched sample based on the propensity score had a greater variance than raw data also confidence interval of OR was wider in the PSM method.[30] It reported similar results to this current study.

PSM has been used successfully by other researchers to aid in creating case-control designs in survey data.[31, 32] In this case-control study, the use of PSM was useful in choosing appropriate matched pairs.

Study Limitations and Suggestions: in this study, other propensity score matching algorithms described in the field of statistics were not examined. Therefore, it is recommended that comparison between other algorithms be performed in future studies.

Many factors contribute to the appearance of CAD in diabetes type 2 patients and only 25% of these are already known.[1] So it might be there are many risk factors as confusing in the relationship between diabetes and coronary artery disease, it is suggested that other studies be performed to match these factors.

Strong points of the study: If the overlap (common support region) is small, that there may not be enough participants in the control group to match all the participants in the case group, then propensity score matching will be no better than any standard form of matching.[16] In this study, there was overlap acceptable.

In a review study in the medical field, only one of the 27 articles had been reported sensitivity analysis.[26] This study reported sensitivity analysis as the last step of matching.

Conclusion

Using different matching algorithms can improve the process of selecting appropriate matched pairs. The PSM method not only guarantees a similar distribution of the confounders in both groups, but it also reduces selection bias. In this study, the caliper algorithm performed better than the nearest neighbor algorithm in selecting appropriate matched pairs.

Acknowledgments

The results of this article are taken from a MSc thesis of Sara Sabbaghian Tousi, with the number of 951748. The authors of the article thank the financial support of the Research Vice-Chancellor of Mashhad University of Medical Sciences.

References

1. Patsouras, A., et al., Screening and Risk Assessment of Coronary Artery Disease in Patients With Type 2 Diabetes: An Updated Review. in vivo, 2019. 33(4): p. 1039-1049.
2. Dehejia, R.H. and S. Wahba, Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics, 2002. 84(1): p. 151-161.
3. Rothman, K.J., S. Greenland, and T.L. Lash, Modern epidemiology. 2008.
4. Allen, A.S. and G.A. Satten, Control for confounding in case-control studies using the stratification score, a retrospective balancing score. American journal of epidemiology, 2011. 173(7): p. 752-760.
5. Cochrane, W. and D. Rubin, Controlling bias in observational studies. Sankyha, 1973. 35(4): p. 417-446.
6. Austin, P.C., A comparison of 12 algorithms for matching on the propensity score. Statistics in medicine, 2014. 33(6): p. 1057-1069.
7. Austin, P.C., Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. The Journal of thoracic and cardiovascular surgery, 2007. 134(5): p. 1128-1135. e3.
8. Austin, P.C., A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in medicine, 2008. 27(12): p. 2037-2049.
9. Austin, P.C., Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical statistics, 2011. 10(2): p. 150-161.

10. Tajfard, M., et al., Anxiety, depression and coronary artery disease among patients undergoing angiography in Ghaem Hospital, Mashhad, Iran. Health, 2014. 6(11): p. 1108.
11. Golpour, P., et al., Comparison of Support Vector Machine, Naïve Bayes and Logistic Regression for Assessing the Necessity for Coronary Angiography. International Journal of Environmental Research and Public Health, 2020. 17(18): p. 6449.
12. Bowers, J., M. Fredrickson, and B. Hansen, RItools: Randomization inference tools. R package version 0.1-11, 2010.
13. Ho, D.E., et al., Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis, 2007. 15(3): p. 199-236.
14. Agresti, A., An introduction to categorical data analysis. 2018: Wiley.
15. Tabesh, H., et al., Prevalence and trend of overweight and obesity among schoolchildren in Ahvaz, Southwest of Iran. Global journal of health science, 2014. 6(2): p. 35.
16. Lechner, M., A note on the common support problem in applied evaluation studies. Annales d'Économie et de Statistique, 2008: p. 217-235.
17. LaLonde, R.J., Evaluating the econometric evaluations of training programs with experimental data. The American economic review, 1986: p. 604-620.
18. Szekér, S. and Á. Vathy-Fogarassy, Weighted nearest neighbours-based control group selection method for observational studies. Plos one, 2020. 15(7): p. e0236531.
19. Rosenbaum, P.R. and D.B. Rubin, Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 1985. 39(1): p. 33-38.
20. Baser, O., Too much ado about propensity score models? Comparing methods of propensity score matching. Value in Health, 2006. 9(6): p. 377-385.
21. Cochran, W.G. and D.B. Rubin, Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A, 1973: p. 417-446.
22. Pan, W. and H. Bai, Propensity score analysis: Fundamentals and developments. 2015: Guilford Publications.
23. Harrell Jr, F.E. and M.C. Dupont, The Hmisc Package. R Package, version, 2006: p. 2.0-0.
24. Rosenbaum, P.R., Observational studies, in Observational studies. 2002, Springer. p. 1-17.
25. Hasegawa, R. and D. Small, Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. Biometrics, 2017. 73(4): p. 1424-1432.
26. Luo, Z., J.C. Gardiner, and C.J. Bradley, Applying propensity score methods in medical research: pitfalls and prospects. Medical Care Research and Review, 2010. 67(5): p. 528-554.
27. Olmos, A. and P. Govindasamy, Propensity scores: a practical introduction using R. Journal of MultiDisciplinary Evaluation, 2015. 11(25): p. 68-88.
28. Austin, P.C., Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. American journal of epidemiology, 2010. 172(9): p. 1092-1097.
29. Austin, P.C., Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations.

Biometrical Journal: Journal of Mathematical Methods in Biosciences, 2009. 51(1): p. 171-184.
30. Pirracchio, R., M. Resche-Rigon, and S. Chevret, Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC medical research methodology, 2012. 12(1): p. 70.
31. Chun, S.-Y., et al., Do long term cancer survivors have better health-promoting behavior than non-cancer populations? Case-control study in Korea. Asian Pac J Cancer Prev, 2015. 16(4): p. 1415-20.
32. Lee, H.S. and J.H. Lee, Vitamin D and urinary incontinence among Korean women: a propensity score-matched analysis from the 2008–2009 Korean National Health and Nutrition Examination Survey. Journal of Korean medical science, 2017. 32(4): p. 661-665.