PLOS ONE

RESEARCH ARTICLE

Weighted nearest neighbours-based control group selection method for observational studies

Szabolcs Szekér[1,2☯]*, **Ágnes Vathy-Fogarassy**[1,2☯]

1 Department of Computer Science and Systems Technology, University of Pannonia, Veszprém, Hungary, **2** Healthcare Business Analytics Research and Development Centre, University of Pannonia, Veszprém, Hungary

☯ These authors contributed equally to this work.
* vathy@dcs.uni-pannon.hu

Citation: Szekér S, Vathy-Fogarassy Á (2020) Weighted nearest neighbours-based control group selection method for observational studies. PLoS ONE 15(7): e0236531. https://doi.org/10.1371/journal.pone.0236531

Editor: Alan D Hutson, Roswell Park Cancer Institute, UNITED STATES

Received: April 14, 2020

Accepted: July 7, 2020

Published: July 23, 2020

Data Availability Statement: Data underlying the study is available on GitHub at https://github.com/vathyfogarassy/WNNEM.

Funding: The publication was supported by Széchenyi 2020 under the EFOP-3.6.1-16-2016-00015. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Although in observational studies, propensity score matching is the most widely used balancing method, it has received much criticism. The main drawback of this method is that the individuals of the case and control groups are paired in the compressed one-dimensional space of propensity scores. In this paper, such a novel multivariate weighted *k*-nearest neighbours-based control group selection method is proposed which can eliminate this disadvantage of propensity score matching. The proposed method pairs the elements of the case and control groups in the original vector space of the covariates and the dissimilarities of the individuals are calculated as the weighted distances of the subjects. The weight factors are calculated from a logistic regression model fitted on the status of treatment assignment. The efficiency of the proposed method was evaluated by Monte Carlo simulations on different datasets. Experimental results show that the proposed Weighted Nearest Neighbours Control Group Selection with Error Minimization method is able to select a more balanced control group than the most widely applied greedy form of the propensity score matching method, especially for individuals characterized with few descriptive features.

Introduction

Case-control studies are widely applied risk analysis methods mainly used in healthcare and social sciences. In retrospective medical studies, the case group typically contains diseased patients, while the control group includes non-diseased individuals, and the study aims to evaluate those risk factors that potentially led to disease occurrence. In the case of prospective studies, typically the effect of a treatment, intervention or other factor is evaluated by comparing the outcome variables of the case (treated) and control (untreated) groups. The popularity of these studies arises from their relatively inexpensive nature, however, the degree of their evidence is lower than randomized trials. The reliability of these studies can be increased by (1) increasing the number of cases included in the study, (2) performing thorough data preparation and data cleaning activities and (3) selecting a proper control group for the case group.

The latter highlights the fact that the primary aim of case-control studies is to compare individuals in the case and control groups. A prerequisite of carrying out appropriate analyses requires that the case and control groups are similar on covariates that predict group membership (e.g. treatment assignment) and affect the examined output. However, fulfilling these requirements is not a trivial task. Many articles have highlighted the importance of the proper implementation of a control group selection method and the effect of unbalanced control groups on the result of analyses [1–5].

Nowadays, propensity score matching (PSM) [6] is the most widely used control group selection method. It is widespread in healthcare analyses [7–9], and is gaining ground in social sciences [10–12] and economics [13–15]. PSM matches the individuals of the case and control groups based on their *propensity score* values, which is the probability of the group (e.g. treatment) assignment conditional on the observed baseline covariates. From a practical viewpoint, various PSM methods exist [16–20] that may differ in terms of selection methodology, the ratio of the treated and untreated individuals or the nature of the selection process. On one hand, individuals can be selected into the control group with or without the replacement of the candidates. A general tendency is to apply PSM with replacement when the population from which the control group is selected is too small. Otherwise, matching without replacement can also be used. On the other hand, the ratio of the case and control groups can also be varied. 1:1 matching is common practice, but in the case of large datasets, other implementations e.g 1:M matching can also be used. Thirdly, the variety of the propensity score-based matching methods also increases by the fact that during the selection of the individuals, greedy or optimal matching can also be applied. In the first case, untreated subject whose propensity score is the closest to the score of a given treated subject is selected and matched. When optimal matching is used, the aim is to minimize the total within-pair difference of the propensity scores, and the pairing is optimized globally [21, 22]. Although, in the literature, different extensions (e.g. radius matching, kernel matching, inverse probability weighting, GPS-CDF method, etc. [23–26]) of the basic PSM methods were also proposed, in everyday practice the greedy 1:1 matching without replacement is the most widely used form of the PSM method.

Despite the popularity of these methods, they have also received much criticism [27–29]. Most of the critical comments point to the possible imbalance between the case group and the control group. It was shown in a recent article [30], that out of 1000 articles using PSM (published between 1983 and 2015), only 6% used any iterative balance checking procedure. In the remaining 94% of the articles, simple 1:1 greedy PSM was applied without any balance checking. King and Nielsen highlighted that PSM is blind to the often large imbalance that can be eliminated by approximating full blocking with other matching methods [30]. Moreover, they pointed out that propensity score matching may increase imbalance even relative to the original data. In a review article published by P. Austin [31] some articles were also identified containing imbalance on the baseline covariates between the case and control groups despite the suitable application of the PSM method. In [32] cardiovascular studies were examined, and authors showed that propensity score methods are not necessarily superior to conventional covariate adjustment.

The main problem with the most widely applied form of the PSM method presumably originates from the application of dimension reduction of the original feature space. Namely, pairing is performed in the 1-dimensional space of the propensity score values, which reduced space might hide the distributions of the original dimensions (features). Covariates that equally affect the probability of individuals belonging to the treated (or diseased) group also affect the value of the propensity score to the same extent. However, the distribution of these variables may be different, and this difference will no longer appear in the 1-dimensional probability values. As long as the matching is performed in the 1-dimensional space of the propensity scores, these differences can not be taken into account during the matching procedure. Publications [33] and [34] also highlighted that matched-pair analysis has to be performed only when matched individuals are highly correlated, but matching subjects having similar propensity scores do not necessarily result in matched subjects with similar covariate values. Based on these considerations, we regard matching in the original n-dimensional vector space or its subspace more suitable than in the 1-dimensional space of propensity scores. The mentioned subspace refers to the covariates which should be included in the propensity score model. Based on recommendations of Austin [35] and Brookhart [36], all variables that affect both the exposure of the group membership and the outcome of the study should always be taken into account.

In this paper, a novel weighted k-nearest neighbour matching method is proposed to select the most suitable individuals into the control group. The proposed method and the most widely used PSM are similar as both methods calculate the probabilities of group membership (e.g. treatment assignment) using a logistic regression (LR) fit. However, PSM pairs the individuals in the reduced 1-dimensional space of propensity scores, the proposed method performs the matching process in the original feature space of covariates utilizing the regression coefficients of the LR model as weight factors of the dimensions. In this way, on the one hand, the proposed method eliminates the main disadvantage of the PSM method, but at the same time retains its advantage by utilizing its regression coefficients as weight factors of the covariates. Monte Carlo simulations show, that the proposed method can overcome the aforementioned shortcomings of the PSM method and may result in a more balanced control group.

In the following, the propensity score matching method and the proposed Weighted Nearest Neighbours Control Group Selection with Error Minimization method (WNNEM) are introduced. Following this, a Monte Carlo study is presented to compare the most widely applied form of the PSM method with the WNNEM method and to highlight the main advantages of the proposed method. To demonstrate the relative utility of the proposed method, some head to head comparisons with other well-known methods (stratified matching, nearest neighbour matching and Mahalanobis metric matching) were also included in our study.

Methods

Matching-based control group selection methods aim to select and pair individuals from a set of potential candidates (X_C) to individuals of the case group (X_T). Individuals $\mathbf{X}_i \in \{X_C \cup X_T\}$ are characterized by n ($n \in \mathbb{N}$) descriptive features (e.g. age, gender, diagnoses) denoted as f_1, $f_2, \ldots, f_n$. Therefore, each subject is denoted as an $\mathbf{X}_i = [x_{if_1}, x_{if_2}, \ldots, x_{if_n}]$ vector of variables, where $i = 1, 2, \ldots, L$ and $L = |X_C \cup X_T|$. The aim of control group selection methods is to select such an $X_{UT} \subset X_C$ control group, that is balanced to the case group, meaning that the distributions of the variables in both sets are similar. Naturally, X_T and X_{UT} must be disjoint sets, that is $X_T \cap X_{UT} = \emptyset$. To ensure this requirement, X_T and X_C must also be disjoint ($X_T \cap X_C = \emptyset$).

Propensity score matching

Propensity score matching refers to matching techniques that are based on propensity scores (PS). Propensity score is the conditional probability of treatment assignment based on the observed baseline covariates:

$$p_i = Pr(z_i = 1|\mathbf{X}_i), \tag{1}$$

where p_i denotes the propensity score for the i-th participant, and $z_i \in \{0, 1\}$ denotes the treatment variable in such a way that $z_i = 0$ refers to the control (e.g untreated) group and $z_i = 1$

refers to the case (e.g. treated) group. Subjects characterized by the same properties have the same propensity scores.

In retrospective observational studies, the true propensity score is unknown and has to be estimated from available data. Usually, it is estimated using a logistic regression model, but other methods have also been examined and used (e.g. recursive partitioning [37], random forests [38], bagging and boosting [39, 40] and neural networks [35, 41]). When the dependent variable is dichotomous, logistic regression is the most commonly used method to estimate the propensity scores. In this case, treatment status is regressed on the observed baseline covariates and propensity scores are estimated by the fitted model. The multiple linear regression function estimated by the logistic regression model can be defined as follows:

$$logit(p) = b_0 + b_1 f_1 + b_2 f_2 + \cdots + b_n f_n \tag{2}$$

where

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{3}$$

and p is the probability of being exposed. Furthermore, the values b_k ($k = 1, 2, \ldots, n$) are the regression coefficients that describe the relative effects of the covariates (f_k) on the status of group (e.g. treatment) assignment. The propensity score estimated by logistic regression is calculated as:

$$p = \frac{e^{(b_0 + b_1 f_1 + b_2 f_2 + \cdots + b_n f_n)}}{1 + e^{(b_0 + b_1 f_1 + b_2 f_2 + \cdots + b_n f_n)}} \tag{4}$$

PSM methods essentially follow the nearest neighbour-based approach. That is, that individual is selected from the candidates pairing whose propensity score is the most similar to the propensity score of the individual to be paired in the case group. If multiple subjects from the candidates have equally close propensity scores to the propensity score of the sample subject, one of those is selected at random. As the greedy method does not contain any restrictions concerning the maximum acceptable difference between the propensity scores of the two matched subjects, practical implementations often take into account a threshold parameter for the selection [42]. Individuals within a certain distance of the propensity scores (*caliper size*) are matched together, and subjects that fall outside this caliper are neglected. Various suggestions have been made for optimal caliper size in the literature [42–44], but usually 0.2 of the standard deviation of the logit of the propensity scores is recommended [42].

Weighted nearest neighbours control group selection with error minimization

As it was described in the Introduction, PSM methods are constantly in the midst of criticism due to the imbalance of the covariates observed in some studies. The disadvantage of the most widely used PSM and all other propensity score-based methods is that they perform the pairing of individuals in the 1-dimensional space of the propensity scores. Additionally, uncertainty is also increased by the fact that propensity scores are estimated and not known a priori. Although earlier some methods have been proposed to pair individuals in the original vector space of the features [45–47], to the best of our knowledge none of them utilizes the result of fitting a logistic regression model during the matching performed in the original vector space. In this section, such a novel weighted nearest neighbours-based control group selection method is proposed in which the relevance of the covariates is estimated by fitting a logistic regression model, furthermore control group selection is performed in the original vector space of the covariates by calculating the weighted distances of the treated individuals and the possible candidates.

The suggested Weighted Nearest Neighbours Control Group Selection with Error Minimization method (WNNEM) considers each subject as an n-dimensional data point in an n-dimensional space, where each covariate (f_k, $k = 1, \ldots, n$) represents a unique dimension. This way, the problem of control group selection can be interpreted as a distance minimization problem. To select a proper control group, we have to identify such individuals from the candidates that lie close to the individuals of the case group. The concept of lying close can be defined in numerous ways. In the case of the proposed method, multivariate matching is performed in which the adjusted odds ratio (OR) values of the fitted multivariable logistic regression model are utilized as weighting factors of the covariates to compute the distances between the individuals.

Before the whole algorithm is presented, two aspects have to be clarified. Firstly, the term distance and secondly the suggested weighting method has to be specified.

Generally, as individuals may be characterized by different types of variables (binary, nominal, ordinal, numerical), the distance calculation method to be applied must be able to handle different data types. Furthermore, as the significance of the covariates may differ, distances have to be calculated separately for each dimension. The third requirement of distance calculation is that the dissimilarity measures with identical values have to express the same degree of dissimilarity.

To fulfil these requirements, the proposed algorithm calculates the differences for each dimension separately and converts all dissimilarity values into the range of [0, 1]. The distance calculation for different data types is performed as follows:

- In the case of binary variables, the simple matching distance is calculated. That is,

$$d_{ij}^{(f)} = \begin{cases} 0 & \text{if } x_{if} = x_{jf} \\ 1 & \text{if } otherwise \end{cases} \tag{5}$$

  where $d_{ij}^{(f)}$ yields the distance of individuals $\mathbf{X}_i \in X_T$ and $\mathbf{X}_j \in X_C$, x_{if} is the value of individual $\mathbf{X}_i$ on binary variable f, and x_{jf} of $\mathbf{X}_j$, respectively.

- In the case of nominal variables, either the simple matching distance presented before (Eq 5) can be calculated or these variables can be coded as a set of binary variables and the distance can be calculated as the normalized distance of binary features, where the normalization constant represents the number of possible values of the nominal variable.

- In the case of numerical variables, the dissimilarity measure can be calculated as the difference of the original values. As the distance calculated in this way depends on the range of the original values, normalization is needed to achieve uniform significance for the same dissimilarities and to make them comparable to the dissimilarity measures calculated on other types of attributes. To fulfil this requirement, min-max normalization must be performed separately for each numerical dimension to map the original values into the range of [0, 1] as follows:

$$x_{if}^{*} = \frac{x_{if} - min_f}{max_f - min_f} \tag{6}$$

  where x_{if} denotes the original value of individual $\mathbf{X}_i$ in the f-th dimension without normalization, min_f represent the minimum and max_f the maximum value measured in the f-th dimension taking into account all individuals from $X_T \cup X_C$, and x_{if}^{*} yields the normalized value of the individual $\mathbf{X}_i$ with regard to the f-th covariate. Subsequently, the distance of individuals $\mathbf{X}_i \in X_T$ and $\mathbf{X}_j \in X_C$ is calculated as the differences of their normalized feature values:

$$d_{ij}^{(f)} = |x_{if}^* - x_{jf}^*| \tag{7}$$

- In the case of ordinal variables, the ordered values have to be coded as ranks and the distance can be calculated as the aforementioned distance of numerical values.

After ensuring that meaning of the dissimilarity values is identical in each dimension, the next step is to weight them according to their relevance to treatment assignment. Previously mentioned, the adjusted odds ratio values of the multivariable logistic regression model fitted on the status of treatment assignment are utilized for this purpose. Generally, the odds ratio is the probability of a characteristic being present divided by the probability of the same characteristic being absent. In our case, the adjusted odds ratio for each independent variable can be obtained by applying the exponential function to the corresponding regression coefficient (b_k) obtained from the multivariable logistic regression model (Eq 2). That is, adjusted odds ratios as the weights of the covariates are calculated as follows:

$$w_k = OR_k = e^{b_k} \tag{8}$$

where w_k denotes the weighting factor of the k-th covariate ($k = 1, 2, \ldots, n$).

The proposed WNNEM method calculates the distances for individuals $\mathbf{X}_i \in X_T$ and $\mathbf{X}_j \in X_C$ the following way:

$$dist(\mathbf{X}_i, \mathbf{X}_j) = \sum_{k=1}^{n} w_k d_{ij}^{(k)} \tag{9}$$

where $d_{ij}^{(k)}$ represents the normalized dissimilarity value of $\mathbf{X}_i$ and $\mathbf{X}_j$ in the k-th dimension, and w_k is the weighting factor of dimension k.

The presented weighted attribute distance is utilized to match the *best pairs* of candidates (X_C) and individuals of the treated group (X_T). Basically, the *best pair* for each $\mathbf{X}_i \in X_T$ is that $\mathbf{X}_j \in X_C$ for which $dist(\mathbf{X}_i, \mathbf{X}_j)$ is minimal. This way, the matching procedure can be regarded an optimization problem, where $\sum_{i,j} dist(\mathbf{X}_i, \mathbf{X}_j)$ has to be minimized.

Our practical experiments show that for 1:1 matching, an adequate solution can be found even without the use of a complex optimization algorithm. The only problem that needs to be handled during optimization is management of the pairing process of those candidates which lie closest to more than one individual from the case group. These candidates are called *candidates in conflict* and formally are defined as follows: $\mathbf{X}_j \in X_C$ is a *candidate in conflict* if $d(\mathbf{X}_i, \mathbf{X}_j)$ is minimal for more than one $\mathbf{X}_i \in X_T$.

For handling these conflicts, the order of the neighbours has to be determined. Let $NN_1(\mathbf{X}_i)$ denote the closest and $NN_2(\mathbf{X}_i)$ the second closest neighbour to individual $\mathbf{X}_i \in X_T$. By definition, $NN_1(\mathbf{X}_i)$ and $NN_2(\mathbf{X}_i)$ are calculated as follows:

$$NN_1(\mathbf{X}_i) = \underset{\mathbf{X}_j \in X_C}{\operatorname{argmin}} \left(dist(\mathbf{X}_i, \mathbf{X}_j) \right) \tag{10}$$

$$NN_2(\mathbf{X}_i) = \underset{\mathbf{X}_j \in X_C - \{NN_1(\mathbf{X}_i)\}}{\operatorname{argmin}} \left(dist(\mathbf{X}_i, \mathbf{X}_j) \right) \tag{11}$$

The design of the conflict-handling method to solve the competition of two individuals was inspired by the Vogel-Korda method: instead of a greedy selection, the second neighbours of the treated individuals are also taken into account. That is, the candidate in conflict is matched to the individual for which the error function is greater. The error function is calculated as the distance of the first and second neighbours of the individuals as follows:

$$E(\mathbf{X}_i) = |dist(\mathbf{X}_i, NN_1(\mathbf{X}_i)) - dist(\mathbf{X}_i, NN_2(\mathbf{X}_i))| \tag{12}$$

In this way, the problem of two competing individuals, $\mathbf{X}_l$ and $\mathbf{X}_m \in X_T$, is solved. In the case of multiple competing individuals, conflicts are handled by dynamic programming. First, the conflict with the largest error is resolved followed by the others in descending order. This principle is applied iteratively until all the conflicts are resolved.

In summary, the steps of the proposed Weighted Nearest Neighbours Control Group Selection with Error Minimization method (WNNEM) are presented by Algorithm 1. It should be noted that although the presented WNNEM algorithm performs 1:1 matching, it can be easily extended for 1:M matching as well. In this case, in Step 3 the set of unpaired elements ($X_{unpaired}$) has to be defined in the way, that each individual of the case group (X_T) should be placed M times into the set $X_{unpaired}$. The other parts of the algorithm do not change.

Algorithm 1: Weighted Nearest Neighbours Control Group Selection with Error Minimization (WNNEM)

Input: X_T case group, X_C set of candidate individuals
Output: X_{UT} control group
1 Perform a logistic regression to estimate w_k weights for all covariates.
2 Normalize X_T and X_C collectively using feature scaling and calculate the $\mathbf{D} = dist(\mathbf{X}_i, \mathbf{X}_j)$ distance matrix for all pairs of individuals of $\mathbf{X}_i \in X_T$ and $\mathbf{X}_j \in X_C$ by Eq 9.
3 Set
  $X_{unpaired} = X_T$
  $X_{UT} = \emptyset$
4 Determine $NN_1(\mathbf{X}_i)$ and $NN_2(\mathbf{X}_i)$ based on the distance matrix $\mathbf{D}$ for all $\mathbf{X}_i \in X_{unpaired}$.
5 Calculate $E(\mathbf{X}_i)$ for all $\mathbf{X}_i \in X_{unpaired}$ by Eq 12.
6 For all $i \in \{arg(X_{unpaired})\}$
  Set $l = \text{argmax}_{\mathbf{X}_i \in X_{unpaired}}(E(\mathbf{X}_i))$
  If $NN_1(\mathbf{X}_l) \notin X_{UT}$:
    $X_{UT} = X_{UT} \cup \{NN_1(\mathbf{X}_l)\}$
    $X_{unpaired} = X_{unpaired} - \{\mathbf{X}_l\}$
    Set $m = arg(NN_1(\mathbf{X}_l))$
    Set $\mathbf{D}(i, m) = \infty$ for all $i \in \{arg(X_T)\}$
7 Repeat *Steps 4 to 6*, till $X_{unpaired} \neq \emptyset$.

Comparison of the PSM and WNNEM methods—Monte Carlo study

Study design. To present the effectiveness of the proposed method, a Monte Carlo simulation-based experimental study was performed. In this study, the quality of the control group arising from the proposed WNNEM method was compared to the quality of the control groups arising from the following matching methods: (i) two greedy PSM methods, (ii) nearest neighbour matching (NNM) [48], (iii) Mahalanobis metric matching (MMM) [47] and (iv) stratified matching (SM) [49]. In the cases using PSM methods, the most widely applied form, namely, the greedy 1:1 propensity score matching performed without replacement of individuals was applied. The only difference between the two forms of the PSM methods applied was the 'caliper size' setting. In the first case (PSM_02), the caliper size was set as 0.2 of the standard deviation of the logit of the propensity scores. In the second case (PSM_DYN), the caliper size in each simulation was determined dynamically and was set at the minimal value for which 1:1 matching could be performed.

As the greedy nature of the PSM, the nearest neighbour matching and the Mahalanobis metric matching makes them sensitive to the order of the candidates, these methods were run 10 times on each generated dataset in such a way, that the matching order was randomized in each experiment. In contrast, when the WNNEM method or stratified matching was applied, the control group selection was performed only once for each of the generated datasets, because of the deterministic nature of these algorithms.

Datasets. For the comparisons, three scenarios of datasets with varying feature characteristics were designed. For each scenario, 100 datasets were generated randomly with the same distribution parameters predefined for the covariates. As a result, each scenario contained 100 different datasets with the same number of individuals. That is, in total, we evaluated the accuracy of the various control group selection methods on 300 datasets. Since the PSM methods, the nearest neighbour matching and the Mahalanobis metric matching ran 10 times on each dataset, the presented research is based on a total of 12600 outcome evaluations.

Scenario I In the first scenario, datasets contained 1000 individuals and correspond to the dataset widely applied in theoretical PSM studies [50]. Individuals were characterized by 10 binary variables, each from a Bernoulli distribution ($x_j \sim \mathrm{B}(0.5), j = 1, \ldots, 10$).

The calculation of the probability of treatment assignment was based on the following logistic regression model:

$$\begin{aligned} \operatorname{logit}(p_{i,treat}) &= b_{0,treat} + \\ & b_L x_{i1} + b_L x_{i2} + b_L x_{i3} + b_M x_{i4} + b_M x_{i5} + \\ & b_M x_{i6} + b_H x_{i7} + b_H x_{i8} + b_{VH} x_{i9} + b_{VH} x_{i10} \end{aligned} \tag{13}$$

where weights b denote low (L), medium (M), high (H) or very high (VH) effect on treatment assignment.

For each subject, a treatment status indicator (Z_i) was generated from a Bernoulli distribution with a subject-specified probability equal to $p_{i,treat}$ ($Z_i \sim \mathrm{B}(p_{i,treat})$). The treated group consisted of subjects where $Z_i = 1$, while subjects where $Z_i = 0$ were assigned to the untreated group (from which the control group was selected). The weights (b_L, b_M, b_H and b_{VH}) were assigned in such a way that approximately 25% of the subjects were treated. The applied weight coefficients were as follows:

- correction for binary: $b_{0,treat} = -1.344090$
- low: $b_L = \log(1.1)$
- medium: $b_M = \log(1.25)$
- high: $b_H = \log(1.5)$
- very high: $b_{VH} = \log(2.1)$

Scenario II The second scenario models such studies in which fewer descriptive variables are available. In this scenario, each individual was characterized by 1 ordinal and 5 binary variables. The ordinal variable represents 5 age groups, while the binary variables may represent, for example, the gender of the subject or various diagnoses. In this scenario, 700 individuals were simulated in each dataset, and the ratio of the candidate subjects to the treated individuals in the 100 datasets was between 2.0 and 3.1.

The assignment of weights to the descriptive variables was as follows: the ordinal variable (x_1) has very high effect, the binary variables have low (x_2), medium ($x_3 - x_5$) and high (x_6) effect on the status of treatment (Eq 14).

$$\begin{aligned} \mathrm{logit}(p_{i,treat}) &= b_{0,treat} + \\ & b_{VH}x_{i1} + b_L x_{i2} + b_M x_{i3} + b_M x_{i4} + b_M x_{i5} + \\ & b_H x_{i6} \end{aligned} \quad (14)$$

The applied weight coefficients, in this case, were the same as in Scenario I.

Scenario III The third scenario is similar to the second one regarding the attributes of individuals and the total number of subjects in each dataset. However, it simulates a more difficult control group selection problem. Although each dataset still contained 700 individuals, the number of treated individuals in the case of the third scenario is higher than in the second one. While in scenario II, the size of the treated group varied between 24.5 and 33.0 percent of the dataset, in the case of the third scenario, it is between 31.7 and 40.7 percent. In other words, the third scenario simulates a more difficult case, where the ratio of the candidate individuals to the treated subjects is lower ($X_C/X_T \approx [1.5, 2.2]$) than in the second scenario ($X_C/X_T \approx [2.0, 3.1]$). Therefore, in this scenario, it is harder to find a proper pair for each treated individual.

To achieve a higher treatment rate, the weight coefficients were modified as follows:

- correction for binary: $b_{0,treat} = -1.344090$
- low: $b_L = \log(1.35)$
- medium: $b_M = \log(1.6)$
- high: $b_H = \log(2.1)$
- very high: $b_{VH} = \log(3.1)$

Methods to evaluate the similarity of the case and control groups. To validate the result of control group selection, the similarity of the case group and the control group has to be evaluated. In our study, comparison was conducted the following ways: (1) comparing the distributions of the covariates and (2) measuring the similarity of the matched pairs.

In the first case, two different performance evaluations were done. On the one hand, the most common balance metric for comparing propensity score methods, namely the standardized mean difference (SMD) [51] was calculated. Additionally, the most commonly applied goodness of fit tests were also calculated. In the case of continuous covariates with normal distribution, Student two-sample *t*-test [52] was applied, and in the case of non-normal continuous covariates the Kolmogorov-Smirnov test [53, 54] was used. In the case of testing the similarity of categorical variables, Chi-square test [55] was applied.

The most common drawback of the previously mentioned methods is that they only examine the similarity of the values of a single covariate. As the Hansen and Bowers imbalance test [56] is applied for more complex evaluation in biomedical studies, this aggregated imbalance measure was also calculated in this study. The main virtue of this measure is that it is an aggregated imbalance test, which allows the imbalance of all covariates to be evaluated simultaneously. The application limit of the test, namely that it can be applied for 1:1 matching without replacement, made it ideal for our study. To extend the complex evaluation for all covariate distributions together, the Distribution Dissimilarity Index (DDI) presented in [57] was also utilized. This measure first compares the disparity between histograms for each covariate separately and calculates the differences of the frequencies of the histogram bins before the differences are totalled and normalized.

As the similarity of the distributions of the covariates does not mean that the paired individuals are also similar, two aggregated pairwise similarity measures, the Nearest Neighbour Index (NNI) and the Global Dissimilarity Index (GDI) [57], were also calculated for each comparison. The common characteristic of these measures is that the overall distance of the case and control groups is calculated by evaluating the pairwise distances of the matched subjects. The NNI examines for each treated subject whether the paired individual is the closest individual in the given population. The GDI evaluates the similarity of the paired elements more precisely. First, it calculates the normalized distances of the paired subjects, then these distances are totalled and normalized for all matched pairs. In this way, the Global Dissimilarity Index expresses the magnitude of the distances of the paired individuals as well.

Results

In this section, the results of the study are presented. For a fair comparison, not only two types of the PSM method were compared to the proposed WNNEM method, but other distance-based approaches (nearest neighbour matching, Mahalanobis metric matching) and the widely used stratified matching were also examined.

Scenario I

As mentioned before, in this scenario each dataset contains 1000 individuals, each characterized by 10 binary covariates.

Table 1 shows the mean value of the NNI, GDI and DDI distance measures with their standard deviations. All presented dissimilarity measures may fall within the range of [0, 1], and the value of zero expresses that the case and control groups are identical. Consequently, the greater the dissimilarity value, the higher the difference of the case and control groups is. The results show that the WNNEM method performed the best in terms of control group selection. All dissimilarity values, both for the paired evaluations (NNI and GDI) and the distribution-based evaluation (DDI), are lower in the case of control groups selected by the WNNEM method than by the other methods.

To evaluate the results of the selected control groups, the similarities of the covariates were also evaluated separately. Firstly, the SMD values were calculated for all covariates and for all matching methods applied. The detailed results are presented in Table 2. It can be seen that all matching methods resulted in well-balanced control groups as all SMD values are less than 0.1. There are only small differences between the values in favour of the distance-based methods. On covariate x_1 the nearest neighbour matching method, on covariates x_2 and x_3 the

Table 1. Dissimilarity measures for Scenario I. The Nearest Neighbour Index (NNI) and the Global Dissimilarity Index (GDI) present the results of the evaluation of the similarity of matched pairs. The Distribution Dissimilarity Index (DDI) evaluates the similarity of the histograms of covariates.

	NNI	GDI	DDI
WNNEM	0.0599±0.0031	0.0613±0.0046	0.0120±0.0033
SM	0.5737±0.0248	0.5737±0.0248	0.5737±0.0248
NNM	0.0649±0.0034	0.0747±0.0057	0.0152±0.0040
MMM	0.0661±0.0036	0.0760±0.0064	0.0150±0.0041
PSM_02	0.3252±0.0290	0.3588±0.0339	0.0664±0.0179
PSM_DYN	0.2875±0.0259	0.3237±0.0327	0.0160±0.0038

https://doi.org/10.1371/journal.pone.0236531.t001

Table 2. SMD values for Scenario I. The table shows the average value of the standardized mean differences and their standard deviations for each covariate separately.

	x_1	x_2	x_3	x_4	x_5
WNNEM	0.034±0.030	0.029±0.027	0.025±0.024	0.022±0.020	0.026±0.021
SM	0.057±0.046	0.057±0.042	0.044±0.036	0.057±0.047	0.057±0.039
NNM	0.024±0.019	0.030±0.022	0.026±0.021	0.028±0.021	0.027±0.021
MMM	0.027±0.026	0.024±0.022	0.023±0.019	0.030±0.024	0.031±0.030
PSM_02	0.039±0.029	0.035±0.024	0.035±0.029	0.033±0.027	0.031±0.024
PSM_DYN	0.034±0.024	0.031±0.023	0.032±0.024	0.028±0.021	0.031±0.023
	x_6	x_7	x_8	x_9	x_{10}
WNNEM	0.023±0.019	0.021±0.018	0.020±0.016	0.022±0.016	0.018±0.014
SM	0.056±0.040	0.059±0.046	0.048±0.032	0.059±0.043	0.065±0.050
NNM	0.027±0.019	0.034±0.027	0.030±0.024	0.033±0.022	0.045±0.030
MMM	0.026±0.024	0.032±0.031	0.027±0.026	0.035±0.033	0.046±0.037
PSM_02	0.035±0.024	0.037±0.027	0.034±0.023	0.036±0.025	0.042±0.029
PSM_DYN	0.029±0.021	0.031±0.025	0.029±0.020	0.028±0.020	0.030±0.022

https://doi.org/10.1371/journal.pone.0236531.t002

Mahalanobis metric matching and on covariates $x_4 - x_{10}$ the WNNEM method resulted in the most balanced control groups. Given that covariates $x_1 - x_3$ were associated only with low weights to the group membership (Eq 13), we can say that the WNNEM method gave the most accurate balancing on the major covariates.

In the next step, the similarity of the covariates for the case and control groups was tested by the Chi-square test. A higher p-value means a more balanced control group in terms of a given covariate. The detailed results are presented as box plots in Fig 1. As can be seen, the median of the p-values for covariates $x_4 - x_{10}$ is the highest in the case of WNNEM method, and for covariates $x_5 - x_{10}$ the interquartile ranges of the WNNEM method is the smallest. Furthermore, the first quartiles from x_4 to x_{10} are also the highest in the case of the WNNEM method.

To summarize the balances of the covariates, the widely used Hansen and Bowers test was also evaluated for all matching algorithms. In the Hansen and Bowers test, covariates are considered poorly balanced if the test value is significant ($p < 0.05$). The higher the p-value, the more similar the case and control groups are. The p-values for all simulations are presented as box plots in Fig 2.

Fig 2 clearly shows that the WNNEM method resulted in the most balanced control groups in terms of the Hansen and Bowers test. Not only the interquartile range is the smallest in the case of the WNNEM method, but the tail of the box as well. Furthermore, only a few outlier simulations can be observed, and, except for one simulation, they all have high p-values. All these facts support the advantages of applying the proposed WNNEM method.

Scenario II

The second scenario simulates such studies in which the age of the patients and another 5 binary parameters are considered as covariates.

The overall statistics of the control group selections are presented in Table 3. It can be seen that the proposed WNNEM method performed better in terms of GDI and DDI measures than the other methods. However, in terms of the Nearest Neighbour Index, the lowest value can be seen at the nearest neighbour matching, but the difference is negligible (0.005). Furthermore, we can observe that distance-based algorithms generally give 1 order of magnitude better results than PSM methods, except the DDI value of the PSM_DYN method. Furthermore,

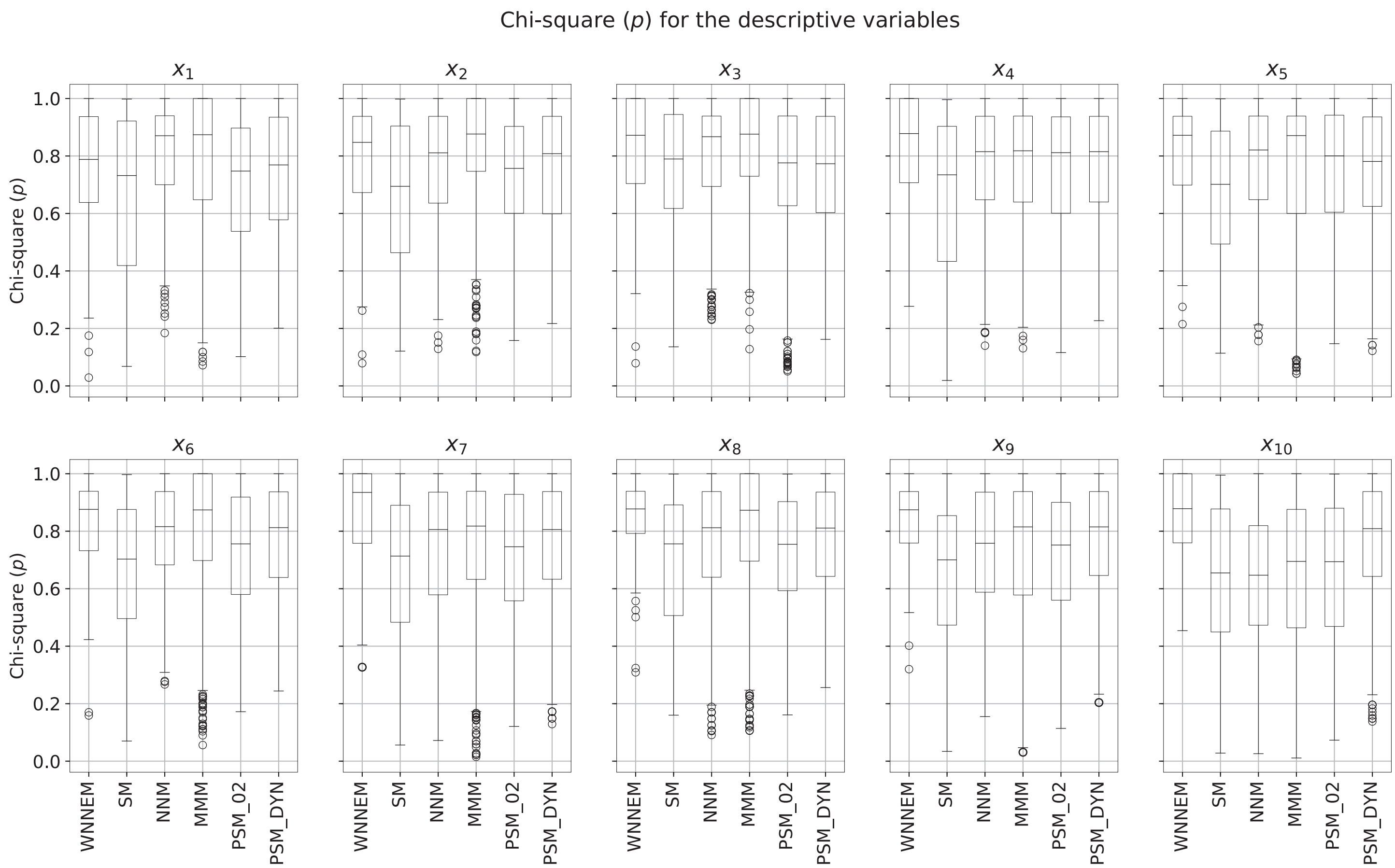

Fig 1. Distribution of covariates in Scenario I. Distribution of the Chi-square *p*-values calculated for each covariate based on all simulations.

https://doi.org/10.1371/journal.pone.0236531.g001

by comparing these results to the results of Scenario I, it can be seen, that in this case, the selected control groups are more similar to the case group.

For further evaluation, the similarity of the covariates was also calculated. Evaluating the SMD measures for the covariates separately, we found that for all covariates and for all methods the average values are below 0.1, so the matched control groups are well balanced on each covariate. The similarity of the case and control groups measured by the Chi-square test is presented in Fig 3. It is important to emphasize that in the case of completely missing boxes, the first, second and third quartiles of the *p*-values were all equal to 1. In the case of partially missing boxes, the median was equal to 1, therefore the third quartile and maximum value were equal. Consistent with the results observed for the NNI, GDI and DDI indices, Fig 3 also emphasizes the better performance of distance-based measures (WNNEM, nearest neighbour matching and Mahalanobis metric matching) for this data set.

Evaluating the Hansen and Bowers test (Fig 4), we can establish that all methods have selected very similar control groups to the case groups. The differences between the boxes are marginal, but the PSM method with a dynamically determined caliper size parameter resulted in less similar control groups more times.

In this example, we have seen that all methods were able to select perfectly balanced control groups. This is due to the lower variability of features of the individuals and the relatively large number of available candidates.

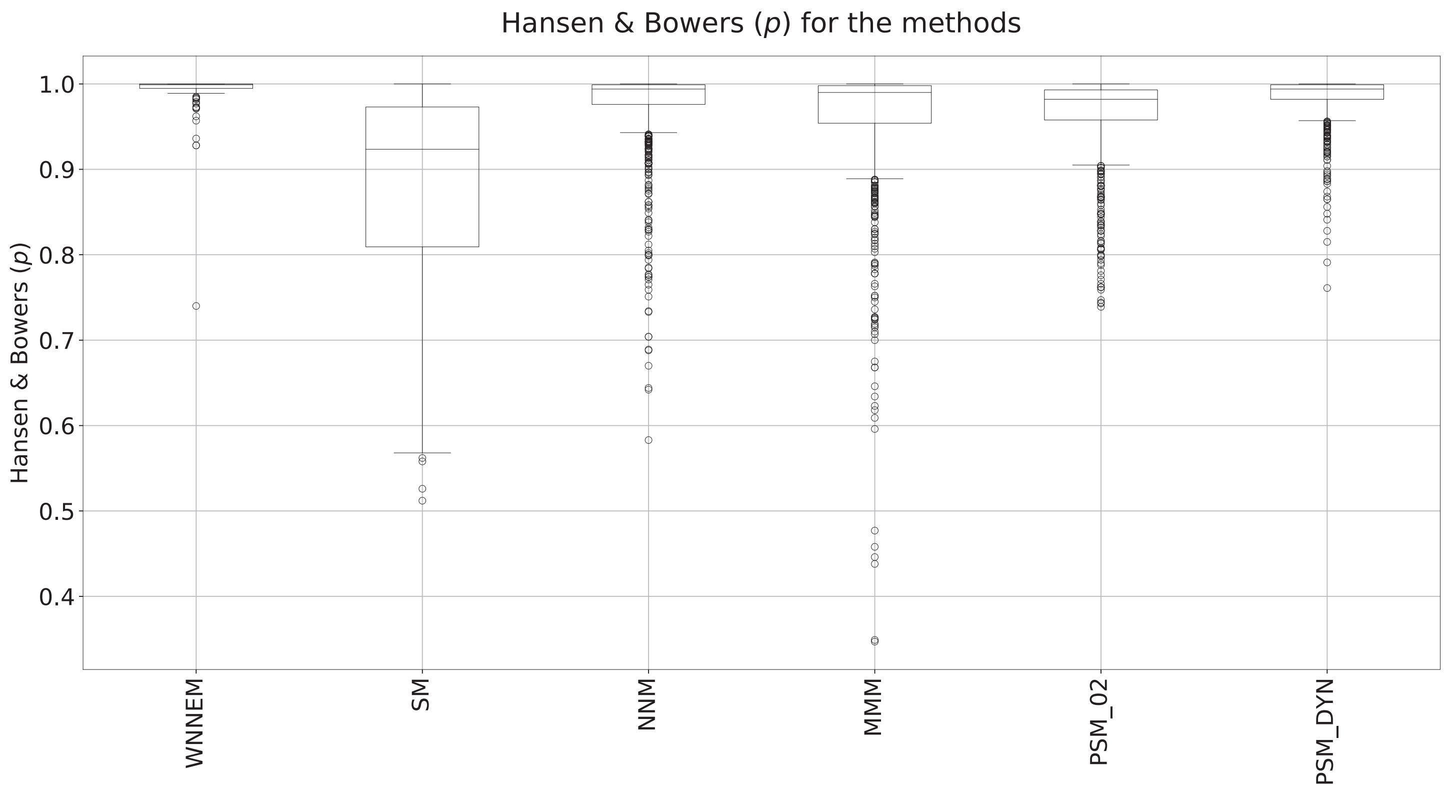

Fig 2. Results of Hansen and Bowers tests in Scenario I. Comparison of the *p*-values of the Hansen and Bowers test for the WNNEM method, the stratified matching, the nearest neighbour matching, the Mahalanobis metric matching and two variants of the PSM method.

https://doi.org/10.1371/journal.pone.0236531.g002

Scenario III

The third scenario is similar to the second one but it simulates a more difficult control group selection problem as the ratio of the candidate individuals to the treated ones is less than in Scenario II. In this scenario, on average, only 1.7 candidate individuals were available per treated person, while in Scenario II this value was 2.5.

In Table 4 the overall dissimilarity measures are presented. By comparing Tables 3 and 4, it can be seen that in the case of the third scenario, it was harder to select fully balanced control groups. However, Table 4 shows that the distance-based methods were still able to select more

Table 3. Dissimilarity measures for Scenario II. The Nearest Neighbour Index (NNI) and the Global Dissimilarity Index (GDI) present the results of the evaluation of the similarity of matched pairs. The Distribution Dissimilarity Index (DDI) evaluates the similarity of the histograms of covariates.

	NNI	GDI	DDI
WNNEM	0.0073±0.0040	0.0069±0.0041	0.0036±0.0028
SM	0.0382±0.0190	0.0382±0.0190	0.0382±0.0190
NNM	0.0065±0.0033	0.0077±0.0040	0.0046±0.0023
MMM	0.0068±0.0035	0.0082±0.0044	0.0041±0.0027
PSM_02	0.0303±0.0169	0.0312±0.0174	0.0251±0.0148
PSM_DYN	0.0162±0.0104	0.0178±0.0119	0.0088±0.0052

https://doi.org/10.1371/journal.pone.0236531.t003

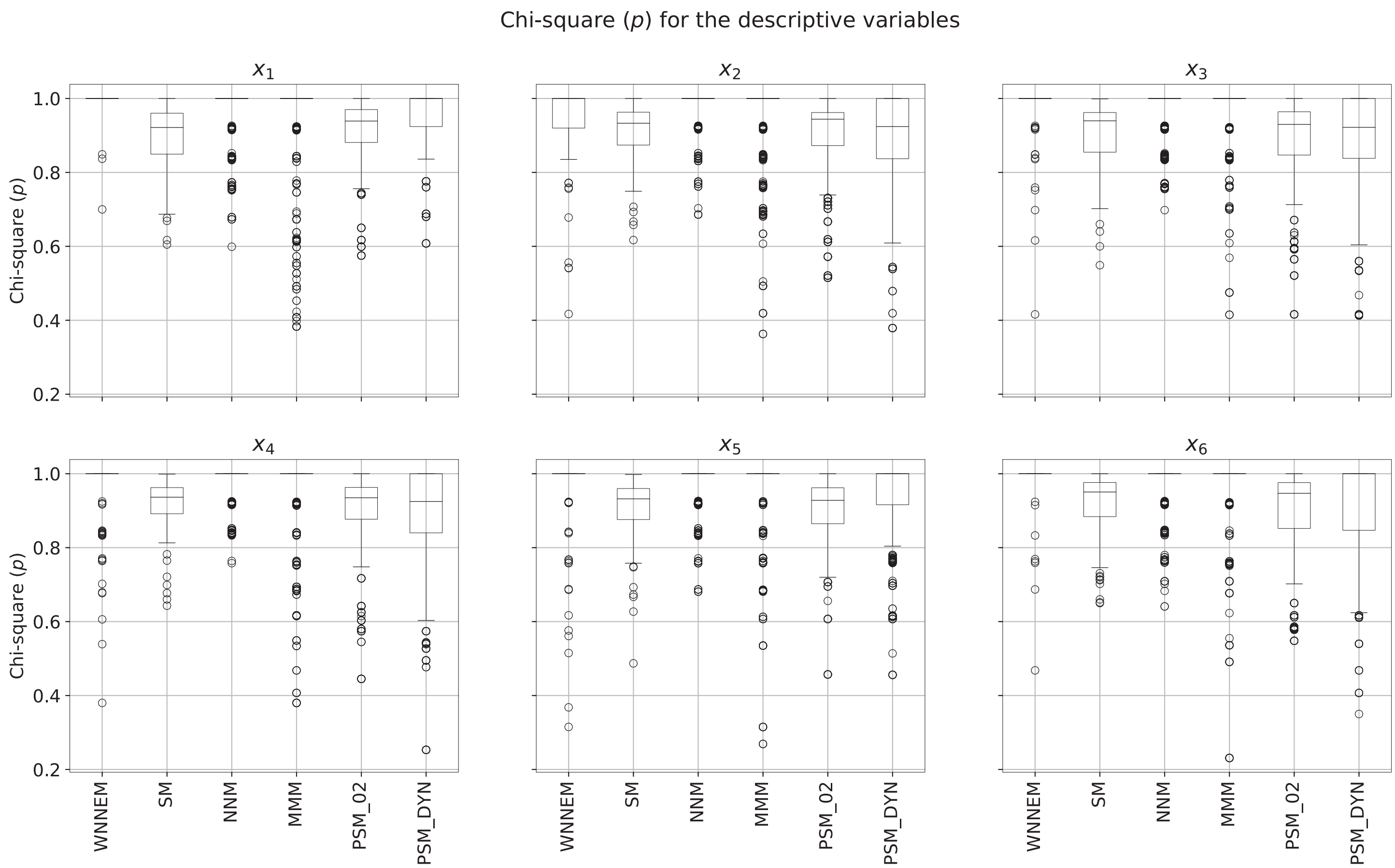

Fig 3. Distribution of covariates in Scenario II. Distribution of the Chi-square *p*-values calculated for each covariate based on all simulations.

https://doi.org/10.1371/journal.pone.0236531.g003

balanced control groups than the greedy PSM methods, and the stratified matching resulted in the worst dissimilarity measures. The nearest neighbour matching and the Mahalanobis metric matching seem to give better results in terms of NNI and GDI measures than the WNNEM method, but the deviation is less than 0.01.

Fig 5 details the covariate imbalances separately. As can be seen, the WNNEM method was able to select better-balanced control groups than the PSM methods and the stratified matching on all covariates. Comparing the WNNEM method to other distance-based methods we can see that the WNNEM method was able to achieve more balanced results on covariates with high (x_6) and very high (x_1) effect on the group assignment (e.g. treatment assignment). The calculated average SMD values were still below 0.1 for all methods.

Fig 6 presents the overall results of the Hansen and Bowers tests. By comparing Fig 4 and Fig 6, it can be seen that in the case of the third scenario, it was harder for each method to select a fully balanced control group. However, Fig 6 shows that the WNNEM method was the most suitable to select a balanced control group. In this way, Fig 6 confirms the ability of the WNNEM method to select better control groups in harder situations. However it should be noted, that the greedy PSM with dynamically determined caliper size also gave good results, but in a few cases the resulting control groups are not well balanced. In the case of using the WNNEM method, it happened only one time.

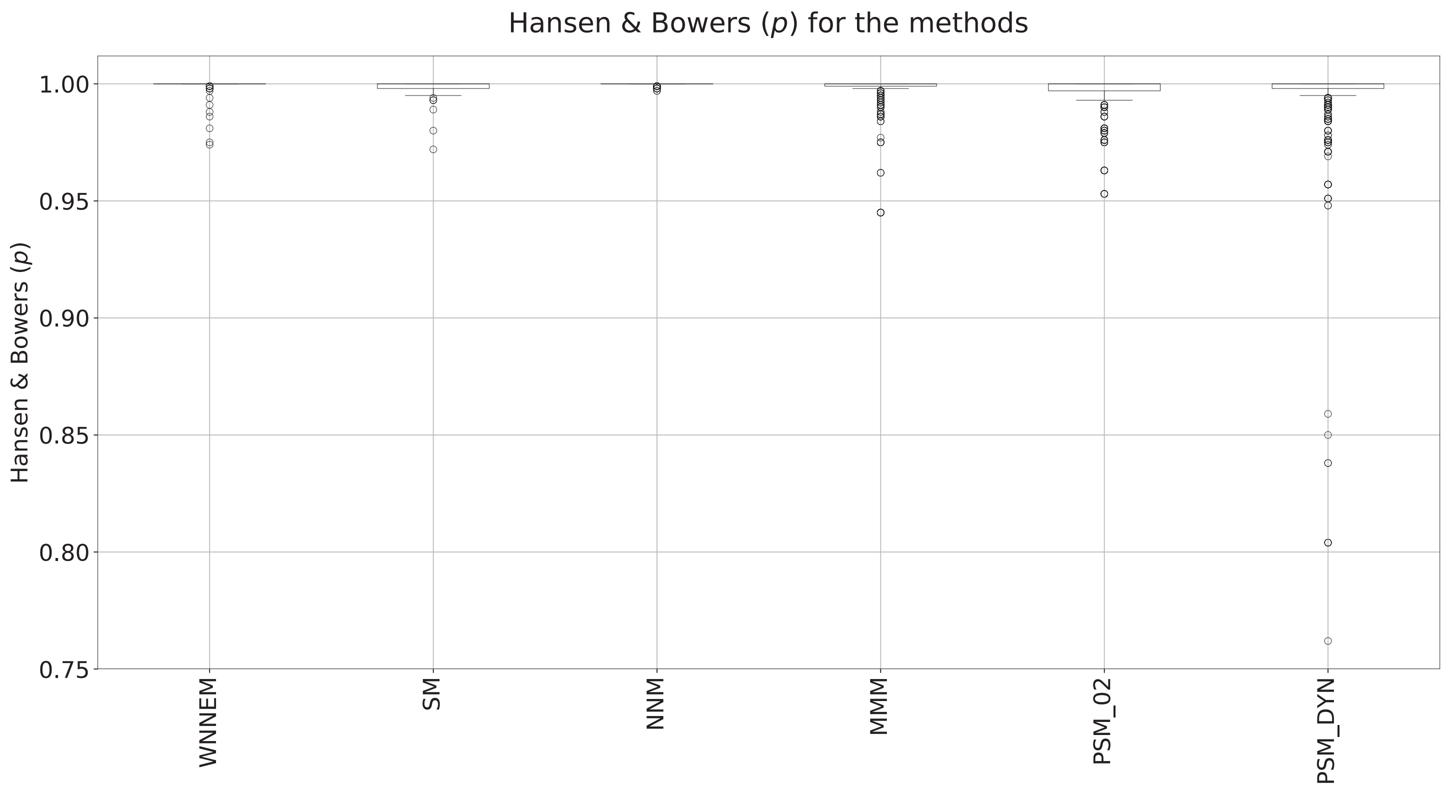

Fig 4. Results of Hansen and Bowers tests in Scenario II. Comparison of the *p*-values of the Hansen and Bowers test for the WNNEM method, the stratified matching, the nearest neighbour matching, the Mahalanobis metric matching and two variants of the PSM method.

https://doi.org/10.1371/journal.pone.0236531.g004

Discussion

Although propensity score matching is the most widely used balancing method, many publications have drawn attention to the disadvantages of using it. Probably, the main problem arises from the fact that the selection of individuals happens in a compressed one-dimensional space of the propensity scores. In this paper, a multivariate control group selection method was proposed, which performs the control group selection in the original vector space of the covariates. The proposed Weighted Nearest Neighbours Control Group Selection with Error Minimization method takes into account the nearest neighbours of the treated individuals and tries to solve the problem of candidates in conflict. Candidates in conflict are those individuals who are the closest to more than one treated subject. Distances between the individuals of

Table 4. Dissimilarity measures for Scenario III. The Nearest Neighbour Index (NNI) and the Global Dissimilarity Index (GDI) present the results of the evaluation of the similarity of matched pairs. The Distribution Dissimilarity Index (DDI) evaluates the similarity of the histograms of covariates.

	NNI	GDI	DDI
WNNEM	0.0304±0.0114	0.0351±0.0165	0.0115±0.0066
SM	0.1151±0.0334	0.1151±0.0334	0.1151±0.0334
NNM	0.0212±0.0065	0.0297±0.0114	0.0144±0.0058
MMM	0.0234±0.0079	0.0321±0.0129	0.0144±0.0067
PSM_02	0.1035±0.0335	0.1093±0.0362	0.0839±0.0304
PSM_DYN	0.0663±0.0255	0.0817±0.0352	0.0187±0.0086

https://doi.org/10.1371/journal.pone.0236531.t004

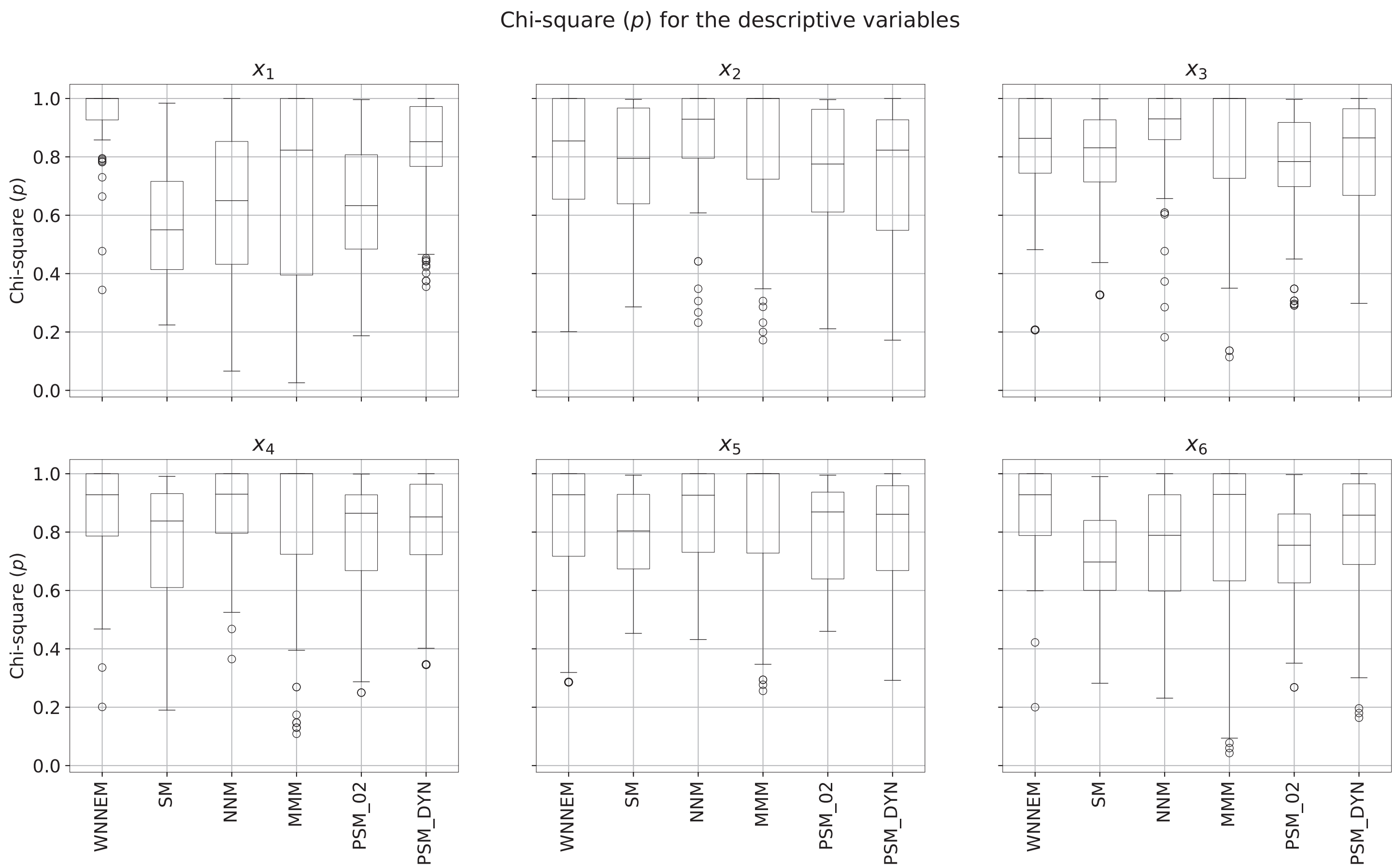

Fig 5. Distribution of covariates in Scenario III. Distribution of the Chi-square *p*-values calculated for each covariate based on all simulations.

https://doi.org/10.1371/journal.pone.0236531.g005

treated and untreated groups are calculated as the weighted sum of the distances of the covariates. As covariates may have different effects on treatment assignment, the proposed WNNEM method weights them according to their relevance. The weighting factors for the covariates are acquired from the logistic regression model fitted on the status of treatment assignment.

Monte Carlo simulations show that in the case of 1:1 matching, the proposed WNNEM method is able to select a more balanced control group than the most widely applied greedy forms of propensity score matching. In this article, we aimed to emphasize the fact that although the most applied forms of PSM can result in a well-balanced control group, a more similar control group can be reached with a simple nearest neighbour-based method proposed in this article. The proposed method is advantageous when individuals are characterized by fewer covariates and the search space is such small that there exist many individuals for selecting as control which are the most similar pairs of more than one treated subject. If more variables are available to characterize individuals, propensity scores describe the exposure of the individuals more precisely, so individuals in conflict may also be less prevalent. As a result, the selected control group is more balanced using the PSM method.

Furthermore, it must be emphasized that the presented form of the WNNEM method solves the problem locally and does not consider further candidates in the selection process. As a result, the proposed WNNEM method does not always select the best balanced control group. This shortcoming of the proposed method can be eliminated by applying a global

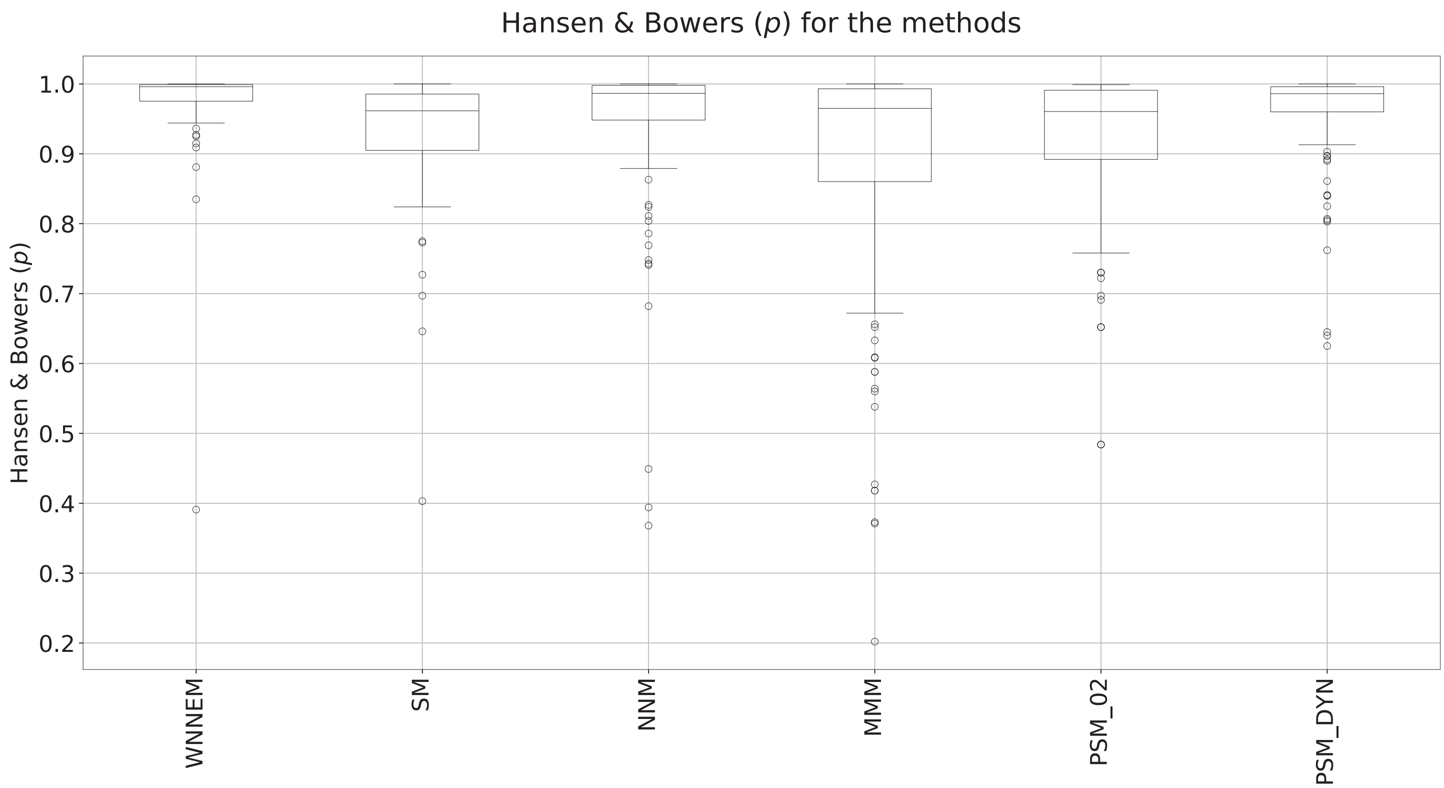

Fig 6. Results of Hansen and Bowers tests in Scenario III. Comparison of the p-values of the Hansen and Bowers test for the WNNEM method, the stratified matching, the nearest neighbour matching, the Mahalanobis metric matching and two variants of the PSM method.

https://doi.org/10.1371/journal.pone.0236531.g006

optimization method, e.g. particle swarm optimization or simulated annealing. In the next step of our research, the advantages of extending the proposed method in this regard will be examined. However, it should be noted that the method proposed in this article is able to select a better control group for small datasets. Moreover, in the presented simple form, this can be achieved without specifying extra parameters. Because of its simplicity, applying the proposed method may be worthwhile, however, further investigations into this topic are necessary. The other limitation of the presented method, that it can only handle covariates with positively associated with the treatment assignment. In our next work, we plan to extend the method to negatively associated covariates as well.

Author Contributions

Conceptualization: Ágnes Vathy-Fogarassy.

Formal analysis: Szabolcs Szekér.

Methodology: Szabolcs Szekér, Ágnes Vathy-Fogarassy.

Software: Szabolcs Szekér.

Supervision: Ágnes Vathy-Fogarassy.

Validation: Ágnes Vathy-Fogarassy.

Visualization: Szabolcs Szekér, Ágnes Vathy-Fogarassy.

Writing – original draft: Szabolcs Szekér, Ágnes Vathy-Fogarassy.

Writing – review & editing: Szabolcs Szekér, Ágnes Vathy-Fogarassy.

References

1. Harris AD, Samore MH, Lipsitch M, Kaye KS, Perencevich E, Carmeli Y. Control-Group Selection Importance in Studies of Antimicrobial Resistance: Examples Applied to Pseudomonas aeruginosa, Enterococci, and Escherichia coli. Clinical Infectious Diseases. 2002; 34(12): 1558–1563. https://doi.org/10.1086/340533
2. Pell GS, Briellmann RS, Chan CHP, Pardoe H, Abbott DF, Jackson GD. Selection of the control group for VBM analysis: influence of covariates, matching and sample size. Neuroimage. 2008; 41(4): 1324–1335. https://doi.org/10.1016/j.neuroimage.2008.02.050
3. Behar P, Teixeira P, Fachel J, Kalil AC. The effect of control group selection in the analysis of risk factors for extended spectrum β-lactamase-producing Klebsiella pneumoniae infections. A prospective controlled study. Journal of Hospital Infection. 2008; 68(2): 123–129. https://doi.org/10.1016/j.jhin.2007.10.022
4. Ripollone JE, Huybrechts KF, Rothman KJ, Ferguson RE, Franklin JM. Implications of the Propensity Score Matching Paradox in Pharmacoepidemiology. American journal of epidemiology. 2018; 187(9): 1951–1961. https://doi.org/10.1093/aje/kwy078
5. Moser P. Out of Control? Managing Baseline Variability in Experimental Studies with Control Groups. In: Bespalov A, Michel M, Steckler T, editors. Good Research Practice in Non-Clinical Pharmacology and Biomedicine. Handbook of Experimental Pharmacology. Vol 257. Springer, Cham; 2019. pp. 101–117.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1): 41–55. https://doi.org/10.1093/biomet/70.1.41
7. Shin SH, Kim SC, Song KB, Hwang DW, Lee JH, Lee D, et al. A comparative study of laparoscopic vs open distal pancreatectomy for left-sided ductal adenocarcinoma: a propensity score-matched analysis. Journal of the American College of Surgeons. 2015; 220(2): 177–185. https://doi.org/10.1016/j.jamcollsurg.2014.10.014 PMID: 25529901
8. Tokuda M, Yamashita S, Matsuo S, Kato M, Sato H, Oseto H, et al. Clinical significance of early recurrence of atrial fibrillation after cryoballoon vs. radiofrequency ablation—A propensity score matched analysis. PLOS ONE. 2019; 14(7): e0219269. https://doi.org/10.1371/journal.pone.0219269 PMID: 31265482
9. Chuang YW, Shih-Ting H, Tung-Min Y, Chi-Yuan L, Mu-Chi C, Cheng-Li L, et al. Acute pancreatitis risk after kidney transplantation: Propensity score matching analysis of a national cohort. PLOS ONE. 2019; 14(9): e0222169. https://doi.org/10.1371/journal.pone.0222169 PMID: 31509567
10. Thoemmes FJ, Kim ES. A systematic review of propensity score methods in the social sciences. Multivariate behavioral research. 2011; 46(1): 90–118. https://doi.org/10.1080/00273171.2011.540475
11. Hwang SH, Cappella E. Rethinking Early Elementary Grade Retention: Examining Long-Term Academic and Psychosocial Outcomes. Journal of Research on Educational Effectiveness. 2018; 11(4): 559–587. https://doi.org/10.1080/19345747.2018.1496500
12. Xu D, Solanki S, Harlow A. Examining the Relationship Between 2-year College Entry and Baccalaureate Aspirants' Academic and Labor Market Outcomes: Impacts, Heterogeneity, and Mechanisms. Research in Higher Education. 2019;
13. Shipman JE, Swanquist QT, Whited RL. Propensity score matching in accounting research. The Accounting Review. 2016; 92(1): 213–244.
14. Cushman DO, De Vita G. Exchange rate regimes and FDI in developing countries: A propensity score matching approach. Journal of International Money and Finance. 2017; 77: 143–163. https://doi.org/10.1016/j.jimonfin.2017.07.018
15. Rosholm M, Mikkelsen MB, Svarer M. Bridging the gap from welfare to education: Propensity score matching evaluation of a bridging intervention. PLOS ONE. 2019; 14(5): e0216200. https://doi.org/10.1371/journal.pone.0216200
16. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics, 2002; 84(1): 151–161. https://doi.org/10.1162/003465302317331982
17. Rubin DB. Matched sampling for causal effects. Cambridge University Press; 2006.
18. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLOS ONE. 2011; 6(3): e18174 https://doi.org/10.1371/journal.pone.0018174 PMID: 21483818
19. Lee W-S. Propensity score matching and variations on the balancing test. Empirical economics. 2013; 44(1): 47–80. https://doi.org/10.1007/s00181-011-0481-0

20. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Statistics in medicine. 2014; 33(6): 1057–1069. https://doi.org/10.1002/sim.6004
21. Rosenbaum PR. Overt bias in observational studies. In: Observational studies. Springer Series in Statistics. Springer; 2002. pp. 71–104.
22. Rosenbaum PR. Design of observational studies. Springer; 2010.
23. Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. Value in Health. 2006; 9(6): 377–385. https://doi.org/10.1111/j.1524-4733.2006.00130.x
24. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research. 2011; 46(3): 399–424. https://doi.org/10.1080/00273171.2011.568786
25. Caliendo M, Kopeinig S. Some Practical Guidance For The Implementation Of Propensity Score Matching. Journal of Economic Surveys. 2008; 22(1): 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x
26. Brown DW, DeSantis SM, Greene TJ, Maroufy V, Yaseen A, Wu H, et al. A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record–derived study. Statistics in Medicine. 2020;
27. Biondi-Zoccai G, Romagnoli E, Agostoni P, Capodanno D, Castagno D, D'Ascenzo F, et al. Are propensity scores really superior to standard multivariable analysis? Contemporary clinical trials. 2011; 32(5): 731–740. https://doi.org/10.1016/j.cct.2011.05.006 PMID: 21616172
28. Pearl J. Remarks on the method of propensity score. Statistics in medicine. 2009; 28(9): 1415–1416. https://doi.org/10.1002/sim.3521
29. Mansournia MA, Jewell NP, Greenland S. Case–control matching: effects, misconceptions, and recommendations. European journal of epidemiology. 2018; 33(1): 5–14. https://doi.org/10.1007/s10654-017-0325-0
30. King G, Nielsen R. Why propensity scores should not be used for matching. Political Analysis. 2019; 27 (4): 435–454. https://doi.org/10.1017/pan.2019.11
31. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in medicine. 2008; 27(12): 2037–2049. https://doi.org/10.1002/sim.3150
32. Elze MC, Gregson J, Baber U, Williamson E, Sartori S, Mehran R, et al. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. Journal of the American College of Cardiology. 2017; 69(3): 345–357. https://doi.org/10.1016/j.jacc.2016.10.060 PMID: 28104076
33. Stuart EA. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics. 2010; 25(1): 1–21. https://doi.org/10.1214/09-STS313
34. Wan F. Matched or unmatched analyses with propensity-score–matched data? Statistics in medicine. 2019; 38(2): 289–300. https://doi.org/10.1002/sim.7976
35. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Statistics in medicine. 2007; 26(4): 734–753. https://doi.org/10.1002/sim.2580
36. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. American journal of epidemiology. 2006; 163(12): 1149–1156. https://doi.org/10.1093/aje/kwj149
37. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and drug safety. 2008; 17(6): 546–555. https://doi.org/10.1002/pds.1555
38. Zhao P, Su X, Ge T, Fan J. Propensity score and proximity matching using random forest. Contemporary clinical trials. 2016; 47: 85–92. https://doi.org/10.1016/j.cct.2015.12.012
39. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological methods. 2004; 9(4): 403–425. https://doi.org/10.1037/1082-989X.9.4.403
40. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Statistics in medicine. 2010; 29(3): 337–346.
41. Cavuto S, Bravi F, Grassi M, Apolone G. Propensity score for the analysis of observational data: an introduction and an illustrative example. Drug Development Research. 2006; 67(3): 208–216. https://doi.org/10.1002/ddr.20079

42. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical statistics. 2011; 10(2): 150–161. https://doi.org/10.1002/pst.433
43. Wang Y, Cai H, Li C, Jiang Z, Wang L, Song J, et al. Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study. PLOS ONE. 2013; 8(12): e81045. https://doi.org/10.1371/journal.pone.0081045 PMID: 24349029
44. Lee M. Matching, regression discontinuity, difference in differences, and beyond. Oxford University Press; 2016.
45. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A. 1973; 35(4): 417–446.
46. Rubin DB. Multivariate matching methods that are equal percent bias reducing, I: Some examples. Biometrics. 1976; 32(1): 109–120. https://doi.org/10.2307/2529342
47. Rubin DB. Bias reduction using Mahalanobis-metric matching. Biometrics. 1980; 36(2): 293–298. https://doi.org/10.2307/2529981
48. Rubin DB. Matching to remove bias in observational studies. Biometrics. 1973; 29: 159–184. https://doi.org/10.2307/2529684
49. Anderson DW, Kish L, Cornell RG. On Stratification, Grouping and Matching. Scandinavian Journal of Statistics. 1980; 7(2): 61–66.
50. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Statistics in medicine. 2011; 30 (11): 1292–1301.
51. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in medicine. 2009; 28(25): 3083–3107. https://doi.org/10.1002/sim.3697
52. Student B. The probable error of a mean. Biometrika. 1908; 6(1): 1–25.
53. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari, 1933; 4: 83–91.
54. Smirnov N. Table for estimating the goodness of fit of empirical distributions. The annals of mathematical statistics. 1948; 19(2): 279–281. https://doi.org/10.1214/aoms/1177730256
55. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1900; 50 (302): 157–175. https://doi.org/10.1080/14786440009463897
56. Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. Statistical Science. 2008; 23(2): 219–236. https://doi.org/10.1214/08-STS254
57. Szekér S, Vathy-Fogarassy Á. How Can the Similarity of the Case and Control Groups be Measured in Case-Control Studies? In Proceedings of IEEE International Work Conference on Bioinspired Intelligence 2019; 33–40.