Health Informatics Meets eHealth
D. Hayn and G. Schreier (Eds.)

doi:10.3233/978-1-61499-759-7-311

Comparison of Control Group Generating Methods

Szabolcs SZEKÉR[a], György FOGARASSY[b] and Ágnes VATHY-FOGARASSY[a,1]
[a]*Department of Computer Science and Systems Technology, University of Pannonia, Hungary*
[b]*State Hospital for Cardiology, Balatonfüred, Hungary*

Abstract. Retrospective studies suffer from drawbacks such as selection bias. As the selection of the control group has a significant impact on the evaluation of the results, it is very important to find the proper method to generate the most appropriate control group. In this paper we suggest two nearest neighbors based control group selection methods that aim to achieve good matching between the individuals of case and control groups. The effectiveness of the proposed methods is evaluated by runtime and accuracy tests and the results are compared to the classical stratified sampling method.

Keywords. Retrospective Studies, Research Design, Control Groups, Matched-Pair Analysis, Statistical Distributions, Sample Size

1. Introduction

In observational medical studies, the goal of the analysis is to identify and evaluate causes of diseases and adverse medical events, or to analyze the effect of specific risk factors for the outcome to be analyzed. The interpretation of results is generally based on a comparative analysis between two independent groups of patients. These groups ideally are very similar to each other, but they differ for a certain characteristic that is in the focus of the study. For example, if we want to evaluate the effect of smoking on lung cancer, then we have to compare the results of smoker patients (case group) with the results of non-smoker individuals (control group).

In cohort studies, subjects are selected by their exposure status and they are followed over a period of time until the outcome of analysis occurs. Because these studies have a temporal framework to assess the causality of the influencing factors, they have the potential to provide the strongest scientific evidence [1]. Cohort studies can be classified as prospective and retrospective studies. Although the methodology of prospective and retrospective cohort studies is fundamentally the same, the study design in these cases is very different because of the different implementation methods [2].

In case of prospective studies, two groups of individuals *(exposed, case group* and *unexposed, control group*) are selected on the bases of factors that are to be examined for possible effects on the outcome. Prospective studies are performed from the present to the future and accordingly, data is collected and recorded during the whole period of follow-up.

[1] Corresponding Author: Ágnes Vathy-Fogarassy, Department of Computer Science and Systems Technology, University of Pannonia, 2. Egyetem Str., 8200 Veszprém, Hungary, E-Mail: vathy@dcs.uni-pannon.hu

In retrospective studies participating individuals can be selected not only on the basis of their exposure status, but they can be also classified as either having some outcome (case group) or lacking it (control group). In retrospective studies, participants are selected in the present, but the examination was carried out in the past. According to this, data about influencing factors and relevant features were measured and recorded also in the past.

Both study methods have some advantages and disadvantages. For example, due to the long time follow-up period, the investigation of prospective cohort studies generally requires many years. In contrast, this is substantially reduced in case of retrospective studies, given that data collection happened in the past and required study time only includes the time of the analysis. However, while in a well-designed prospective study the selection of the participating individuals into case and control groups can be designed and performed in a predetermined manner, in case of retrospective studies this cannot be done in such a way. Generally, the selection of the case group can be carried out based on the study aims, but the determination of the control group has difficulties and it raises many questions [3].

Wacholder and his coworkers determined three principles for the selection of the control group in case-control studies [4]. In this work the comparability is defined as (1) all comparisons must be made within the study base, (2) the comparison of the effects of the levels of exposure on the outcome must not be distorted by the effects of other factors and (3) any errors in measurement of exposure must be non-differential between cases and controls (comparable accuracy). Ensuring these principles is not a simple task. If control group selection is not appropriate, the third principle of comparability is unsatisfied. In retrospective studies, the investigator has limited control over data collection and the maximal size of the analyzable population is predetermined, so control group selection principles may be damaged and significant biases may affect the selection of controls.

In the literature a lot of different methods have been proposed to select control groups for case-control studies. In the simplest case, the individuals of the control group are generally selected by stratified sampling (SS) [5]. In these cases, strata are defined based on the predictive variables. This methodology is proper if the size of the set of possible candidates is large enough and the distribution in each stratum is corresponsive. Otherwise, the selection of individuals for control cannot be performed properly.

Using balancing scores such as propensity score (PS) offers an alternative solution for selecting proper individuals into the case group. Propensity score is the probability of treatment assignment conditional on observed baseline characteristics. There are numerous ways of utilizing PS in control group selection, such as matching, stratification, inverse probability with PS as weight or covariate adjustment [6]. The most popular may be propensity score matching (PSM) which gives a solution to the abovementioned problem by matching treated and untreated subjects who share a similar value of PS. The weakness of PSM comes from its attempt to approximate a completely randomized experiment. This property makes PSM uniquely blind to often large imbalances that can be eliminated by approximating full blocking with other matching methods. Moreover, for adequately balanced data PSM approximates random matching which increase imbalance even relative to original data [7].

In this paper we present two novel nearest neighbor based control selection methods to solve the problem of control group selection. In contrast to propensity score matching, the suggested methods do not consider the influencing effect of the observed covariates. Our aim was to develop such control group selection methods that can ensure the same

distributions of all measured variables in the control group as in the case group. Stratified sampling considers all measured variables, not only the influencing ones. Thus the efficiency of the proposed methods was compared to the well-known stratification-based method and to the basic nearest neighbor based selection method. Our tests show, that the suggested methods may outperform the efficiency of the classical methods.

The structure of the paper is the following. Section 2 introduces the well-known and newly developed methods. In Section 3 the results of the comparative tests are presented. Finally, Section 4 concludes the paper.

2. Methods

Retrospective approach infers various constraining factors, given, that data collection took place in the past. The number of available variables for the analysis is limited and therefore it is difficult to take into account the effect of possible confounding variables. This makes control selection a nontrivial problem. Before we present some solutions for this problem, let us introduce the following notations.

Given a *population P* characterized by variables $f_1, f_2, \dots, f_n (n \in \mathbb{N})$ (for example, age, gender, etc…). *P* denotes the group of individuals of the retrospective study that are investigated. Denote P_A the *case group*, which is a subset of P ($P_A \subset P$). Our goal is to determine such a $P_B \subset P$ ($P_A \cap P_B = \emptyset$) *control group*, in which the distribution of $f_1, f_2, \dots, f_n$ holds as in the case group P_A. $P_A \cap P_B = \emptyset$ means, that the individuals of the control group should be different from the individuals of the case group. The elements of the control group are selected from the set $P_C = P - P_A$, which we call as *candidate subpopulation*.

2.1. Sampling-based Control Group Selection Methods

Traditional methods of control group selection often utilize simple randomized sampling or stratified sampling. As random sampling does not consider the similarity of the case group and the control group, in aspect of the investigation variables, we do not deal with this method in detail.

A more sophisticated method is stratified sampling. Stratified sampling divides the members of the population into homogeneous subgroups (strata) before sampling, reducing sampling error. Every element in the population must be assigned to one and only one stratum based on their values of variables $f_1, f_2, \dots, f_n$. The elements of the control group are selected from these strata based on the frequencies of the individuals with these values in the case group. The main problem of stratified sampling lies within the strata. On one hand, if the number of variables and their recorded values are numerous, we have to generate exponentially large number of strata. On the other hand, if the size of a stratum is inadequate (that is the stratum contains not enough individual element from the candidate subpopulation), we cannot select enough individuals from that stratum into the control group. Consequently, if constraints on the size of the control group are unsatisfied, the result will be biased.

2.2. Nearest Neighbor-based Control Group Selection Methods

Another way to approach the problem might be nearest neighbor-based control group formation. The *k* nearest neighbor based clustering method may offer a better solution for the problem. Let us consider each element of the population as an *n*-dimensional data point in the *n*-dimensional space, where each relevant variable $(f_1, f_2, \dots, f_n)$ represents a unique dimension. In this case, the problem of control group selection is translated into a distance-minimization problem. To find proper people into the control group, we have to select those individuals from the candidate subpopulation which lie close to the individuals of the case group. Namely, if individuals are close to each other in the *n*-dimensional space, they are similar to each other as well. The concept of closeness can be defined different ways. In our research the distance of the individuals was calculated as the weighted distance of different type of variables [6]. Naturally, this method does not guarantee, that matched individuals will coincide on all the features, but the degree of the similarity can be determined in the function of distance.

In the simplest case, to select the individuals into the control group we have to find the closest element from the candidate set for each individual in the case group. More formally, an adequate control group can be achieved by calculating the distance between each $\mathbf{x_i} = (x_1, x_2, \dots, x_n) \in P_A, (i = 1, 2, \dots, N_A)$ and $\mathbf{y_j} = (y_1, y_2, \dots, y_n) \in P_C$ $(j = 1, 2, \dots, N_C)$, and selecting those $\mathbf{y_j}$ for each $\mathbf{x_i}$, where $d(\mathbf{x_i}, \mathbf{y_j})$ distance is minimal. The notation N_A represents the number of individuals in the case group and N_C stands for the number of the candidate individuals. The main steps of this basic nearest neighbor based control group selection algorithm can be summarized as follows:

2.3. Algorithm 1: Nearest Neighbor based Control Group Selection method (NNCS)

Step 1: Calculate the distance for each pair of individuals from the case group and the candidate group.
Step 2: Select the nearest neighbor from the candidate subpopulation for each individual in the case group into the control group.

However, selecting people like this violates the uniqueness of the elements of matched control group. An individual from the control group can belong to more than one patient in the case group, which violates the aforementioned size constraint. For this reason, we have developed two extended nearest neighbor based algorithms, which ensure the uniqueness of elements in the control group.

The *Extended Nearest Neighbor based Control Group Selection Method* (ENNCS) ensures the uniqueness of the control group by eliminating conflicts. A *conflict occurs*, when an individual in the candidate subpopulation $(\mathbf{y_j} \in P_C)$ is selected as the nearest neighbor for more than one individual in the case group $(\mathbf{x_{i_1}}, \mathbf{x_{i_2}}, \dots \in P_A)$. In the extended version of the nearest neighbor based algorithm in such a scenario, $\mathbf{y_j}$ is assigned to that $\mathbf{x_{i_k}}$ element in the case group for which $d(\mathbf{x_{i_k}}, \mathbf{y_j})$ is minimal. To select the proper pair for the unmatched individuals in the case group the next nearest neighbor is selected. This iterative process repeats until for each individual in the case group a unique element of the candidate set is selected. The algorithm can be summarized as follows:

2.4. Algorithm 2: Extended Nearest Neighbor based Control Group Selection method (ENNCS)

Step 1: Calculate the distance for each pair of individuals from the case group and the candidate group.
Step 2: Select the nearest neighbor from the candidate subpopulation for each individual in the case group into the control group and delete them from the candidate group. Yield all individuals in the case group as matched element.
Step 3: If an element from the candidate subpopulation was selected as the nearest neighbor for more than one person in the case group, then assign this candidate element as the matched pair to the closest individual in the case group. All other individuals in the case group, to which this candidate element was the closest pair, yield as unmatched.
Step 4: Delete the matched elements from the case group, and repeat from Step 2.

ENNCS eliminates the problem of conflicting nearest neighbors, however, it does not take into account the distance of the second neighbors of the case elements. This implies that the aforementioned distance may be excessively large. By avoiding these biases the accuracy of the ENNCS algorithm can further improved.

The *Nearest Neighbor based Control Group Selection Method with Error Minimization algorithm* (NNCSE) utilizes the same notion of conflicts as ENNCS but eliminates them in a different manner. The elimination is based on the following error function:

$$Err(\mathbf{x_i}, \mathbf{y_j}) = \frac{1}{|d(\mathbf{x_i}, \mathbf{y_j})) - d(\mathbf{x_i}, nn(\mathbf{x_i})^{(2)})|} \quad (1)$$

where $\mathbf{x_i}$ is an individual from the case group, with $\mathbf{y_j}$ as nearest neighbor and $nn(\mathbf{x_i})^{(2)}$ the second nearest neighbor of $\mathbf{x_i}$ from the candidate group. If an $\mathbf{y_j}$ individual from the candidate group is the nearest neighbor for more than one person in the case group, then NNCSE algorithm matches this individual to that $\mathbf{x_i}$ for which $Err(\mathbf{x_i}, \mathbf{y_j})$ is minimal. In short, a selected candidate is assigned to that individual in the case group for which the next closest neighbor is farther. By doing so, overall error becomes minimal.

2.5. Algorithm 3: Nearest Neighbor based Control Group Selection method with Error Minimization (NNCSE)

Step 1: Calculate the distance for each pair of individuals from the case group and the candidate group.
Step 2: Select the nearest neighbor from the candidate subpopulation for each individual in the case group into the control group and delete them from the candidate group. Yield all individuals in the case group as matched element.
Step 3: If an element ($\mathbf{y_j}$) from the candidate subpopulation was selected as the nearest neighbor for more than one person in the case group, then assign this candidate element as the matched pair to the individual in the case group ($\mathbf{x_i}$) for which $Err(\mathbf{x_i}, \mathbf{y_j})$ is minimal. All other individuals in the case group, to which this candidate element was the closest pair, yield as unmatched.
Step 4: Delete the matched elements from the case group, and repeat from Step 2.

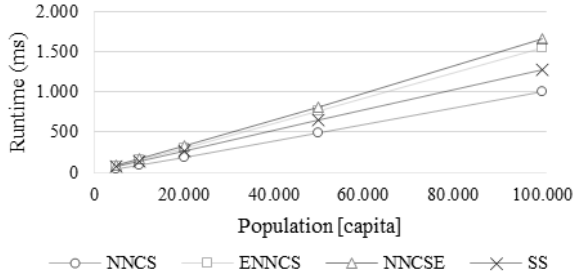

Figure 1 Runtime results of different control group selection methods (size of case group: 1000 people; desired size of control group: 1000 people**)**

3. Results

To evaluate the effectiveness of the aforementioned methods we have performed accuracy and runtime tests. These tests were performed on the anonymized dataset of cancer patients selected from the Hungarian financial health care database. Our main objective was to determine and compare the runtime and accuracy of the stratification-based method (SS) and the nearest neighbor based methods (NNCS, ENNCS, NNCSE). All four algorithms were implemented in Python using the NumPy and Panda libraries and test results were evaluated on a computer with 8Gbs of RAM and a 2 core 4th-gen Intel i5 CPU with Hyperthreading, running a Windows 10 64bit operating system.

During the tests, multiple scenarios were realized to give us a widespread understanding of how the described methods behave under different conditions in point of runtime and preciseness. All scenarios consisted of a case group with 1000 patients suffering from colorectal cancer and the aim was to create a control group of 1000 patients. Available population ranged from 5000 to 200000; namely 5000, 10000, 20000, 50000, 100000 and 200000. Larger populations contained all individuals of the smaller populations. The following features characterized each patient: age, gender and the presence of some diseases, such as insulin-dependent diabetes mellitus, non-insulin-dependent diabetes mellitus, essential hypertension, old myocardial infarction and the presence of other malignancy.

3.1. Runtime

Runtime tests were performed to evaluate the time required to generate the control group. As a reference, we used the stratified sampling (SS) method. In case of SS, the required

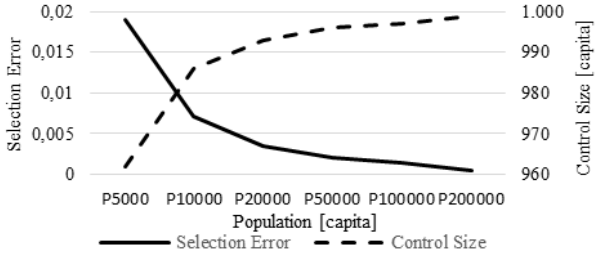

Figure 2 Selection error and the size of the control group in case of SS with increasing population size (size of case group: 1000 people; desired size of control group: 1000 people)

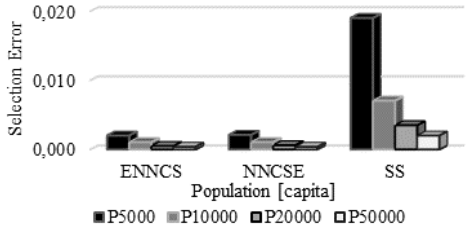

Figure 3 Selection error for SS, ENNCS and NNCSE with increasing size of population from 5000 to 50000 (P5000, P10000, P20000, P50000)

times for the scenarios ordered by the size of the population were as follows: 76.86ms, 140.48ms, 267.61ms, 648.75ms, and 1280.78ms. We can see, that the runtime is a linear function of the size of the population.

Of the nearest neighbors-based methods (NNCS) was the fastest algorithm. It's roughly 17% faster than SS, 38% faster than ENNCS and 45% faster than NNCSE. ENNCS and NNCSE, by ensuring that the size of the selected control group is adequate, are the slowest methods. However, it is important to notice, that while being the slowest ones, their runtimes are still under 2 seconds in the worst case as well. Runtime results can be seen in Figure 1.

3.2. Precision

The precision of the methods was evaluated based on a selection error (SE) that is the bias of the resulted control group and takes into account the number of individuals in a strata. SE was computed as follows:

$$SE(P_A, P_B) = \frac{\Sigma_i \Sigma_j \left| |f_{ij}|_{P_A} - |f_{ij}|_{P_B} \right| - n|N_A - N_B|}{2N_A} \quad (2)$$

where $|f_{ij}|_{P_A}$ yields the number of the individuals characterized by the j-th value of the i-th feature in the case group, and analogously $|f_{ij}|_{P_B}$ in the control group. If the distribution of the generated control group is the same as the distribution of the case group, then selection error is 0.

As mentioned before, the precision of stratified sampling is in relation with the size of the population. Figure 2 shows, that by increasing the size of the population, the selection error is decreasing, as expected. If the size of a stratum is inadequate, we cannot select enough individuals from that stratum. For example, if the population contained 5000 individuals, the SS algorithm was only able to select 962 people into the control group from that population. Therefore, the requirement formulated for the size of the control group could not be met.

The basic NNCS method is not appropriate to generate control groups as the selection errors in this case may be very high. For example, this algorithm has selected only 526 people into the control group from the population containing 5000 people, and this value did not even exceed 600 at best with a selection error ranging between 0.311 and 0.223

In the case of ENNCS and NNCSE algorithms, the selection errors are lower than the selection error of the SS method, even at a smaller population (see Figure 3). As these algorithms always guarantee the required size of the control group, the selection error is

arising only from some biases. With a population of 5000 individuals, the selection error was around 0.002, which is around the value at SS with a population of 200000 people. This value further decreases by increasing the size of the population, reaching a selection error of only 0.0004 in the case of population of 50000. These results justify our initial thoughts that control group generating with the improved versions of nearest neighbor based selection can be upgraded to a more effective level.

4. Conclusion

In this article, we have proposed two nearest neighbors based control group generating methods. These methods place patients in an *n*-dimensional space according to the number of the characterizing variables (*n*). The individuals are selected into the control group from a candidate subpopulation in the function of the distance from the patients in the case group. While the proposed Extended Nearest Neighbor based Control Group Selection Method (ENNCS) takes into account only the first neighbors of the individuals, the Nearest Neighbor based Control Group Selection Method with Error Minimization algorithm (NNCSE) looks further and optimizes the selection error locally. The efficiency of the proposed ENNCS and NNCSE algorithms was compared to the classical stratified sampling and to the basic nearest neighbor selection method. Results show, that ENNCS and NNCSE algorithms offer a reasonable alternative to stratified sampling and the basic version of the nearest neighbor selection method is not appropriate for control group generation. However, stratified sampling is a very fast algorithm, it does not guarantee the predefined size of the control group. In contrast, ENNCS and NNCSE algorithms can achieve it at almost the same speed, but with better precision even for a smaller population. Our future goals include the evaluation of EENCS and NNCSE by comparing them to a PSM based implementation.

Acknowledgement

This publication has been supported by the Hungarian Government through the project VKSZ 12-1-2013-0012 - Competitiveness Grant: Development of the Analytic Healthcare Quality User Information (AHQUI) framework.

References

[1] Everitt BS, Palmer CR., Encyclopaedic Companion to Medical Statistics, Hodder Arnold, London, 2005.
[2] Song, Jae W., Kevin C. Chung, Observational Studies: Cohort and Case-Control Studies, Plastic and reconstructive surgery 126(6) (2010), 2234–2242.
[3] Koepsell, Thomas D., and Noel S. Weiss, Epidemiologic methods: studying the occurrence of illness, Oxford University Press (UK), 2014.
[4] Sholom Wacholder, Joseph K. McLaughlin, Debra T. Silverman, and Jack S. Mandel, Selection of Controls in Case-Control Studies, Americal Journal of Epidemology 135(9) (1992), 1019-1028.
[5] Jewell NP., Least squares regression with data arising from stratified samples of the dependent variable, Biometrika 72 (1985), 11-21.
[6] Peter C. Austin, An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Onservational Studies, Multivariate Behav Res 46(3) (2011), 399-424.
[7] Gary King, Richard Nielsen, Why Propensity Scores Should Not Be Used for Matching. 1st ed. Available at: j.mp/psnot (Accessed: 2 January 2017)