Algoritmo de pareamento para geração de base homogênea para estudo de caso-controle

Matching algorithm for homogeneous base generation for case-control study

Algoritmo de coincidencia para generación de bases homogéneas para estudio de casos y controles

Alexandre da Costa Sena[2], Alexandre Ribeiro Fernandes Azevedo[1], Gabriel Pereira Mendes[1], Karla Figueiredo[2], Luís Cristóvão Moraes Sobrino Pôrto[3]

1 Msc/PhD student, Matematical and Statistics Institute, State University of Rio de Janeiro, Rio de Janeiro, Brazil.
2 PhD/Associate Professor, Matematical and Statistics Institute, State University of Rio de Janeiro, Rio de Janeiro, Brazil.
3 PhD/Associate Professor, Piquet Carneiro University Polyclinic (PPC), State University of Rio de Janeiro, Rio de Janeiro, Brazil

Autor correspondente: (Prof. Dr.) Alexandre da Costa Sena
E-mail: asena@ime.uerj.br

Abstract

Objective: generate balanced control bases in relation to the case base for a case-control study (CCS) and, therefore, minimize selection bias. Method: implementation and evaluation of a matching algorithm that generates homogeneous control bases, based on the characteristics of the case base. Results: the algorithm is capable of producing a homogeneous control file, with the same characteristics as the control base through three different real data sets. Furthermore, the algorithm is generic and easy to use and can be adopted for any case-control study. Conclusion: this algorithm has the potential to greatly help the scientific community, increasing the reliability of research carried out through case-control studies, generating a homogeneous database, helping to avoid selection bias.

Keywords: Matching algorithm; Case Control Study; Big Data

DOI: 10.59681/2175-4411.v16.iEspecial.2024.1293

Resumo

Objetivo: gerar bases de controle balanceadas em relação a base de casos para estudo caso-controle (ECC) e, com isso, minimizar o viés de seleção. Método: implementação e avaliação de um algoritmo de pareamento para gerar bases de controle homogêneas, a partir das características da base de casos. Resultados: o algoritmo é capaz de produzir um arquivo de controle homogêneo, com as mesmas características da base de controle, através de três conjuntos de dados reais distintos. Ainda, o algoritmo é genérico e fácil de usar, podendo ser adotado para qualquer estudo de caso. Conclusão: o algoritmo tem potencial para ajudar a comunidade científica, aumentando a confiabilidade das pesquisas realizadas através de estudo caso-controle, gerando uma base de dados homogênea, ajudando a evitar o viés de seleção.

Descritores: Algoritmo de pareamento; Estudo Caso-Controle; Big Data

Resumen

Objetivo: generar bases de control equilibradas en relación con la base de casos para un estudio de casos y controles y, por tanto, minimizar el sesgo de selección. Método: implementación y evaluación de un algoritmo de emparejamiento que genera bases de control homogéneas, en función de las características de la base de casos**.** Resultados: el algoritmo es capaz de producir un archivo de control homogéneo, con las mismas características que la base de control utilizando tres conjuntos de datos reales diferentes. Además, el algoritmo es genérico y fácil de usar y puede adoptarse para cualquier estudio de caso. Conclusión: este algoritmo tiene el potencial de ayudar enormemente a la comunidad científica, aumentando la confiabilidad de las investigaciones realizadas a través de estudios de casos y controles, generando una base de datos homogénea.

Descriptores: Algoritmo de coincidencia; Estudio de casos y controles; Grandes datos

Introduction

A case-control study (CCS) is a type of observational study widely used to investigate factors associated with disease. This type of study begins with a group of individuals with an outcome (disease or condition) of interest, called cases. Then, based on the characteristics of the individuals in the case group, the researcher builds a second group, called control, composed of people with the same characteristics, but who do not have the outcome of interest. Finally, the researcher analyzes possible factors that led to a certain outcome [1]. CCS is a relatively cost-effective and efficient study design to establish evidence to support future prospective studies or to develop new hypotheses, as it is performed retrospectively with available patient data [2].

When designing a case-control study, the researcher must find an appropriate control group. Ideally, the case group and the control group should have the same characteristics, such as age, sex, general health status, among other factors [1]. Furthermore, choosing controls according to the characteristics of the case base leads to a balanced number of cases and controls at all levels of the selected corresponding variables. This balance can reduce variance in the parameters of interest, which improves statistical efficiency [3]. The selection of an appropriate control group can be difficult, being subject to selection bias, which is the situation where cases and controls differ systematically [4].

On the other hand, due to great technological advances, there is currently an immense amount of data that could be used for case-control studies. In particular, in the healthcare sector, big data sources include hospital records, patient medical records, medical test results, and devices that are part of the Internet of Things. Biomedical research also generates a significant amount of data relevant to public health [5].

In this context, the aim of this work is to propose and evaluate a matching algorithm for generating balanced databases for case-control studies, facilitating the researcher's task, especially when the amount of data is large, minimizing the problem of selection bias. The algorithm allows the user to define the quantity and which attributes should be matched. This algorithm has the potential to greatly help the scientific community, allowing the creation of a homogeneous control base in relation to cases based on characteristics identified as risk exposure factors, helping to avoid selection bias.

The evaluation of the proposed matching algorithm, using three different databases, showed that it is capable of producing a fully balanced control file, especially for the highest priority fields. Furthermore, when comparing with two other algorithms (No Filter and Random) that do not use any priority, it is possible to verify the efficiency of the proposed algorithm. Finally, the evaluation also showed the importance of defining the priority order of attributes so that the algorithm generates a balanced control base for all characteristics of interest.

No matching algorithm for generate a control based on the characteristics of the case base was found in the literature. On the other hand, there are many works that present methods for analyzing matching case-control studies (i.e. analyze if the case and control base that will be used in the study is balanced), where the predominant analysis method is conditional logistic regression. This method provides a conditional estimate of the probability ratio of being ill, given the exposure of interest and the co-variables (3). More recently, the work in (6) showed that, by pre-processing the data, any linear and non-linear classification algorithm can be used to adequately analyze the data generated by the matched case-control study. It is important to highlight that, unlike the algorithm proposed in this work, which aims to generate balanced control files in relation to the case base, this type of algorithm is used to evaluate whether the generated base is balanced or not.

Differently from the matching algorithm presented in this work, which is generic and scalable and can be used for any case study, an initial version specific for analyzing the prevalence of HLA alleles in patients with COVID-19 was presented in (7). In this version, only attributes from the respective CCS could be chosen, using a single case study-specific priority order. Furthermore, the comparison to be performed in each field was also fixed, limiting its use to a single CCS.

Methods

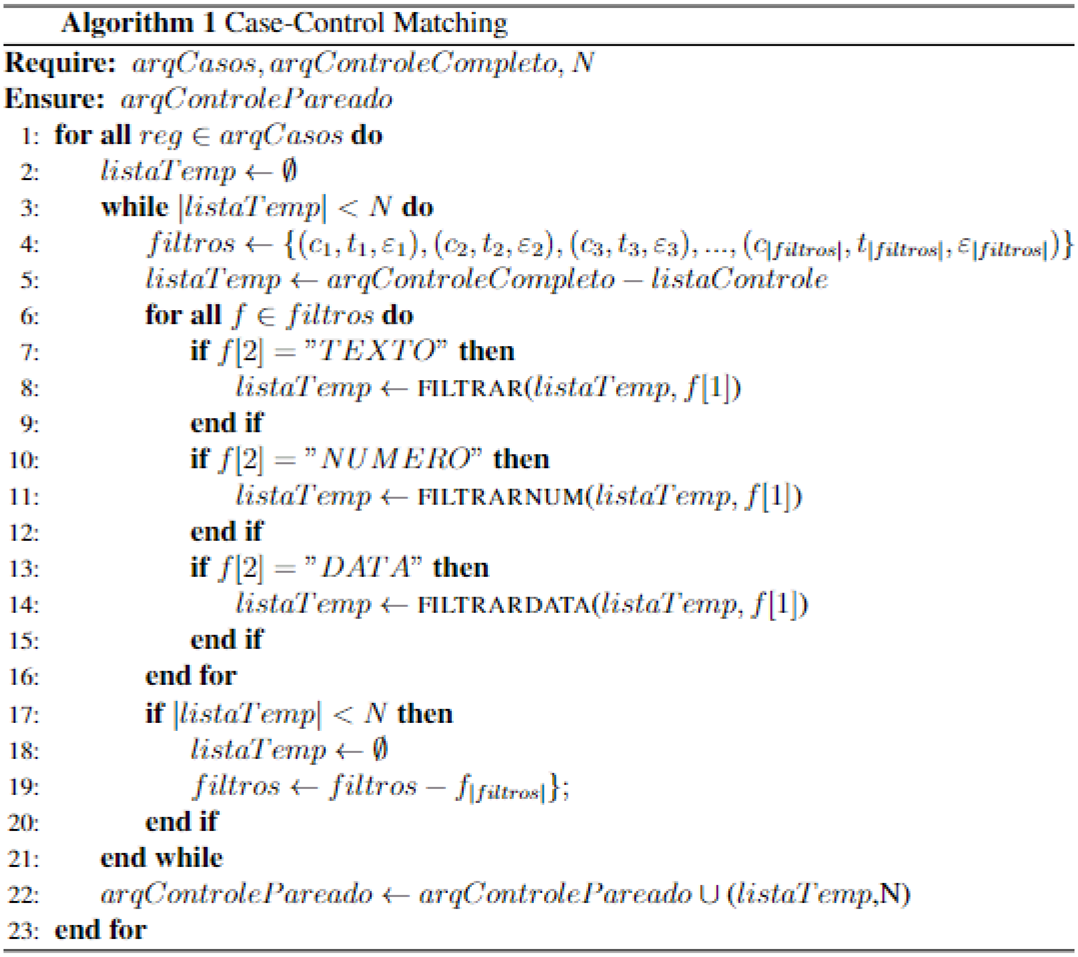
To achieve a balanced case-control study (CCS) and reduce possible selection bias, this work proposes and implements a matching algorithm that chooses the control base, based on characteristics (e.g. gender, ethnicity, age, region, among others) of the case base records. This section describes in detail the proposed and implemented algorithm. Initially, the user must define 3 algorithm parameters: the attributes to be balanced and their order of priority, the type of comparison to be performed in each attribute and the number of records to be chosen in the control base for each record in the case base. These three parameters are explained below.

The choice of attributes to be balanced and their priority order is made through a graphical user interface (GUI). The first field (attribute) chosen will have the highest priority, the second field chosen will have the second highest priority, and so on. Then, for each of the chosen fields, the user must define the type of comparison (filter) to be performed. Currently, the algorithm has three types of comparisons: TEXT, NUMBER and DATE. A field defined as being of type TEXT must be exactly the same for the algorithm to consider it to be matched. For example, when comparing the SEX field in the case base with the SEX field in the control base, it will only be considered equal (matched) if the values are exactly the same, including accents and whether the letters are in upper or lower case. In other words, if the value of the SEX field in the case base is "Female", the same value must be found in the control base to be matched.

In turn, for a field defined as NUMBER, the user must enter a value ε that will define the range for the matching. For example, if the value of the case base AGE field is 23 and ε = 5, then the control base record will be considered matched if the value of the AGE field is in the range {18 ≤ AGE ≤ 28}. Still, for a field defined as DATE, the operation is the same as the treatment given to fields of the NUMBER type, that is, a value must be defined for ε that will define the initial value and the final value for the date to be considered matched. For example, if the value of the DATE field in the case base is 15/06/1981 and ε = 10, then the record in the control base will be considered paired if the value of the DATE field is in the range {05/06/1981 ≤ DATE ≤ 25/06/1981}.

Finally, the user must choose the value of N that defines the dimension in the case:control relationship (1:N). For example, if the user defines the value N=3 (1:3), then for each record in the case base, 3 records from the control base will be matched. Next, in Figure 1, the matching algorithm for the Case-Control Study is presented.

Figure 1 – Case-Control Matching Algorithm



Algorithm 1 Case-Control Matching

```
Require: $arqCasos, arqControleCompleto, N$
Ensure: $arqControlePareado$
 1: for all $reg \in arqCasos$ do
 2:     $listaTemp \leftarrow \emptyset$
 3:     while $|listaTemp| < N$ do
 4:         $filtros \leftarrow \{(c_1, t_1, \varepsilon_1), (c_2, t_2, \varepsilon_2), (c_3, t_3, \varepsilon_3), ..., (c_{|filtros|}, t_{|filtros|}, \varepsilon_{|filtros|})\}$
 5:         $listaTemp \leftarrow arqControleCompleto - listaControle$
 6:         for all $f \in filtros$ do
 7:             if $f[2] = "TEXTO"$ then
 8:                 $listaTemp \leftarrow$ FILTRAR($listaTemp, f[1]$)
 9:             end if
10:             if $f[2] = "NUMERO"$ then
11:                 $listaTemp \leftarrow$ FILTRARNUM($listaTemp, f[1]$)
12:             end if
13:             if $f[2] = "DATA"$ then
14:                 $listaTemp \leftarrow$ FILTRARDATA($listaTemp, f[1]$)
15:             end if
16:         end for
17:         if $|listaTemp| < N$ then
18:             $listaTemp \leftarrow \emptyset$
19:             $filtros \leftarrow filtros - f_{|filtros|}\};$
20:         end if
21:     end while
22:     $arqControlePareado \leftarrow arqControlePareado \cup (listaTemp$,N)
23: end for
```

The algorithm receives as input the case (*arqCasos*) and control (*arqControleCompleto*) files, in addition to the quantity N that defines the 1:N dimension of the matching. The *arqControleCompleto* file contains all candidate control records to be part of the study. The output of the algorithm is the matched control file (*arqControlePareado*) which contains exactly the number of records in the case base multiplied by N.

The matching for each record, reg, in the case base (*arqCasos*) is performed from line 1 to 23 of the algorithm. In line 2, the variable *listaTemp*, which will temporarily store the control base records that were matched, is initialized. The loop in line 3 of the algorithm controls whether the N records from the case base with the filters applied were found. The filters (comparisons) to be applied are defined in line 4 of the algorithm. Each filter is a tuple that contains the field to be filtered, field type and the ε value that defines the range, if they are NUMBER or DATE fields. For example, the filter f=(c1,t1,ε1) could define that the first filter to be applied is the field c1=SEX and that the field type is t1 = TEXT (as it is a field of type TEXT, ε1 is not used). It is important to highlight that the order of the filters defines the priority order to be adopted by the algorithm. In this example, the sex field filter will have the highest priority because it is in the first position.

In turn, the *listaTemp* variable is initialized with all records from the complete control base that were not used in the matching (line 5). The loop in line 6 will apply all the filters chosen by the user (variable *filtros*). At each step of this loop, the filter with the highest priority at the moment is applied according to its type (i.e. TEXT, NUMBER or DATE). For example, assuming that the highest priority filter is SEX and that this field in the control base is FEMALE, then the FILTER function will select only records in the temp list variable where the SEX field is FEMALE. Furthermore, assuming that the second highest priority filter is AGE, then the FILTRARNUM function will select only the records from the *listaTemp* variable (that is, records with the SEX field = FEMALE) where the AGE field is within the Range AGE_{cases} - ε to AGE_{cases} + ε. The loop continues until all filters are applied. If, at the end of the loop, the number of records in *listaTemp* is greater than N, that is, enough records were found in the control base with characteristics of the case base record, then the first record in the case base is matched. Thus, the first N records of the *listaTemp* are added to the variable *arqControlePareado*, which at the end of the algorithm will contain the entire matched control base (line 22). However, if the number of records in the *listaTemp* is less than N, then it is necessary to restart the filtering process for that record with one less filter (line 17 to 20). Note that the filter removed is always backwards (lower priority). In this case, for this record, there will be no balancing for all filters.

The entire process described so far is repeated for each of the records in the case base (main loop of line 1 of the algorithm). In the end, the control dataset balanced according to the characteristics of the case base records can be accessed in the file *arqControlePareado*.

Results and Discussion

To evaluate the proposed matching algorithm, three different sets of real data, called **Case 1**, **Case 2** and **Case 3,** were used in this work. Each of these cases corresponds to a specific condition of interest. Furthermore, all data used in this study was received with an anonymized coding of records, in line with the General Law for the Protection of Personal Data (LGPD), in accordance with ethical principles outlined in the Declaration of Helsinki. In order to compare the results produced by the proposed matching algorithm, two other algorithms were created. The algorithm called **NoFilter** is basically the same algorithm presented, but without applying any filter. In other words, it chooses the first N records available in the complete control base for each record in the case base without applying any filters. On the other hand, the algorithm called **Random** selects N records for each record from the case base randomly. The algorithm proposed in this work is called **Matched**, as it applies all the filters chosen by the user. Furthermore, to verify the performance of the algorithm according to the increase in the size of N, each of the 3 algorithms was executed with values 2, 3 and 5.

To evaluate the matching results, the Epi Info statistical software designed by the CDC (Centers for Disease Control and Prevention) for the global community of physicians and public health researchers was used [(8)]. The results of the matchings obtained for each of the 3 case-control studies evaluated in this work are presented below.

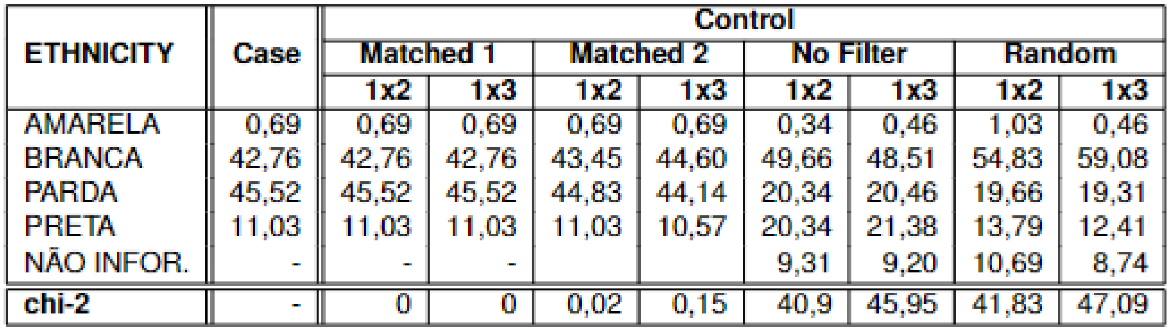
Case 1

The first case study is the one with the largest number of records (145 records in case base and 5006 records in the control base). In this study, two different matchings were generated to show how the priority order of attributes (fields) can drastically influence the result. Thus, exceptionally for this study, only values 2 and 3 were used for N. The only reason for not carrying out the experiment with N=5 is to improve the visualization of the results that compare these two matchings. The relevant characteristics (attributes) for this study were ETHNICITY, SEX, AGE and STATE. The attributes ETHNICITY, SEX and STATE are of type TEXT, while the AGE field is of type NUMBER. The value of ε was 5.

That is, the algorithm initially tried to pair ages from the control file that were in the interval (CaseAge - 5) ≤ ControlAge ≤ (CaseAge + 5).

Thus, as explained previously, for this case study, two matchings were carried out with different order of priorities: i) **Matched 1** with priority order ETHNICITY > SEX > AGE > STATE. ii) **Matched 2** with priority order STATE > ETHNICITY > SEX > AGE. The frequency (in percentage) for each of the ethnicities present in the case and control files generated by the algorithms, as well as the chi-square value are presented in Figure 2.

Figure 2 – Frequency in percentage (%) for the values of the ETHNICITY attribute (Case 1).



ETHNICITY	Case	Control							
		Matched 1		Matched 2		No Filter		Random	
		1x2	1x3	1x2	1x3	1x2	1x3	1x2	1x3
AMARELA	0,69	0,69	0,69	0,69	0,69	0,34	0,46	1,03	0,46
BRANCA	42,76	42,76	42,76	43,45	44,60	49,66	48,51	54,83	59,08
PARDA	45,52	45,52	45,52	44,83	44,14	20,34	20,46	19,66	19,31
PRETA	11,03	11,03	11,03	11,03	10,57	20,34	21,38	13,79	12,41
NÃO INFOR.	-	-	-			9,31	9,20	10,69	8,74
chi-2	-	0	0	0,02	0,15	40,9	45,95	41,83	47,09

When comparing the frequency values for each of the ethnicities in the control files of the matchings generated with the two different priority orders with the values from the case base, one can see the perfect distribution produced by **Matched 1** execution, as well as a distribution close to perfection by **Matched 2**. That is, the change in priority of the ETHNICITY attribute (field) from the first (**Matched 1**) to the second (**Matched 2**) position did not harm the balance of this attribute, despite the slight worsening. On the other hand, when we analyze the values produced by the other two algorithms (**No Filter** and **Random**), it is possible to notice a high degree of unbalance.

To compare two categorical variables and check whether they are homogeneous with each other, the chi-square test was used. In this case, the ETHNICITY variable from the case base was compared with the ETHNICITY variables from the control base generated by the algorithm. The values for chi-2 in the table clearly show that the two executions with different orderings of the proposed algorithm (**Matched** 1 and **Matched** 2) produced the most homogeneous results, since the lower the chi-2 value the better. In fact,

Matched 1 was perfect (value 0), with a 100% probability of being homogeneous and **Matched** 2 was very close to perfection with chi-2 values 0.02 and 0.15, with probabilities 99.9% and 98, 5% of being homogeneous. On the other hand, the chi-2 values for the other two algorithms were very high, with a 0% (zero) probability of being homogeneous.

Differently, the result for the origin STATE field was very poor with the last priority (**Matched** 1) and perfect when executed with the highest priority (**Matched** 2). The results can be viewed in Figure 3.

The frequencies of the control base of the STATE attribute generated with the lowest priority (**Matched** 1) were very far from the frequencies of the case base, including choosing records from states that did not belong to the case base. Furthermore, when analyzing the chi-2, the values were very high, indicating an unbalanced sample. On the other hand, when the control file was generated with the highest priority for the STATE attribute (**Matched** 2), the frequencies of the case and control bases were identical and, consequently, with chi-2 equal to 0 and a probability of 100% of the bases being homogeneous. Finally, as expected, the results produced by the other two algorithms produced unbalanced bases.

Figure 3 – Frequency in percentage (%) for the values of the STATE attribute (Case 1).

STATE	Case	Control							
		Matched 1		Matched 2		No Filter		Random	
		1x2	1x3	1x2	1x3	1x2	1x3	1x2	1x3
BA	0,69	2,07	1,61	0,69	0,69	-	0,23	0,69	0,92
CE	2,07	-	-	2,07	2,07	-	-	0,34	-
ES	1,38	0,69	1,15	1,38	1,38	1,03	0,92	0,69	0,23
MA	0,69	-	0,46	0,69	0,69	-	-	-	0,23
MG	2,76	1,38	1,61	2,76	2,76	2,07	2,07	1,38	2,53
PA	0,69	-	0,23	0,69	0,69	0,34	0,23	-	-
PB	2,07	1,03	0,69	2,07	2,07	-	0,46	1,03	0,23
PE	1,38	0,34	0,23	1,38	1,38	0,69	0,69	-	0,23
PR	0,69	0,34	0,23	0,69	0,69	-	0,23	-	0,69
RJ	86,90	90,34	90,57	86,90	86,90	91,72	91,72	93,45	91,49
RN	0,69	0,34	-	0,69	0,69	-	-	0,34	-
AL	-	0,34	0,23	-	-	-	-	0,34	-
AM	-	0,34	0,23	-	-	-	-	-	-
AP	-	0,34	0,23	-	-	0,34	0,23	-	0,23
DF	-	-	-	-	-	-	-	0,34	0,46
GO	-	0,34	0,23	-	-	0,34	0,23	-	-
PI	-	0,34	0,23	-	-	0,34	0,46	0,34	0,46
RO	-	0,69	0,46	-	-	0,34	0,23	-	0,46
SC	-	-	-	-	-	0,34	0,23	-	-
SE	-	-	-	-	-	0,69	0,46	-	-
SP	-	1,03	1,38	-	-	1,72	1,61	1,03	1,84
chi-2	-	20,5	24,5	0	0	27,3	26,6	19,1	31,9

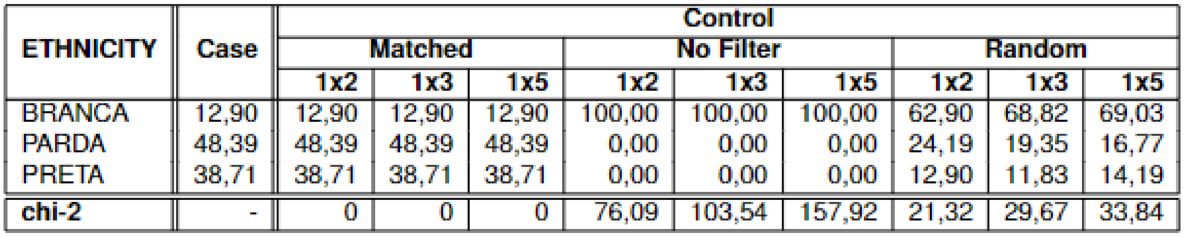
Regarding the SEX attribute, the change from the second highest priority (**Matched** 1) to the third highest (**Matched** 2), caused the control base to go from being perfectly balanced to a slight increase in the frequency of women. The last attribute to be analyzed was AGE. While the arithmetic mean of the ages of the control base was very far from the mean of the case base, the median was very close for the two matchings (**Matched** 1 and 2). The main reason for this difference in the arithmetic mean was the greater number of records with an age much older than the ages in the case base (e.g. the oldest age in the case base is 80 years, while the oldest age in the complete control base is 55 years).

Case 2

The second case study was composed of 31 records in the case base and 1466 records in the complete control base. The relevant characteristics (attributes) for this study were ETHNICITY, SEX and AGE. The attributes ETHNICITY and SEX are of type TEXT, while the AGE field is of type NUMBER, with value of ε = 5.

Figure 4 presents the frequency (in percentage) for each of the ethnicities present in the case file and control files generated by the 3 algorithms, as well as the chi-square value. When comparing the results, one can see the perfect distribution produced by the **Matched** algorithm. On the other hand, when we analyze the values produced by the other two algorithms, we notice a much higher percentage, where the **Random** algorithm presented better results than the **No Filter** one.

Figure 4 – Frequency in percentage (%) for the values of the ETHNICITY attribute (Case 2).



ETHNICITY	Case	Control								
		Matched			No Filter			Random		
		1x2	1x3	1x5	1x2	1x3	1x5	1x2	1x3	1x5
BRANCA	12,90	12,90	12,90	12,90	100,00	100,00	100,00	62,90	68,82	69,03
PARDA	48,39	48,39	48,39	48,39	0,00	0,00	0,00	24,19	19,35	16,77
PRETA	38,71	38,71	38,71	38,71	0,00	0,00	0,00	12,90	11,83	14,19
chi-2	-	0	0	0	76,09	103,54	157,92	21,32	29,67	33,84

The values for chi-2 in the table clearly show that the **Matched** algorithm produced the most homogeneous results, since the lower the chi-2 value the better. In fact, in this case it was perfect (value 0), with a 100% probability of being homogeneous. In turn, the chi-2 values for the other two algorithms were very high (i.e. for values above 5 the test is considered not valid), with a 0% (zero) probability of being homogeneous.

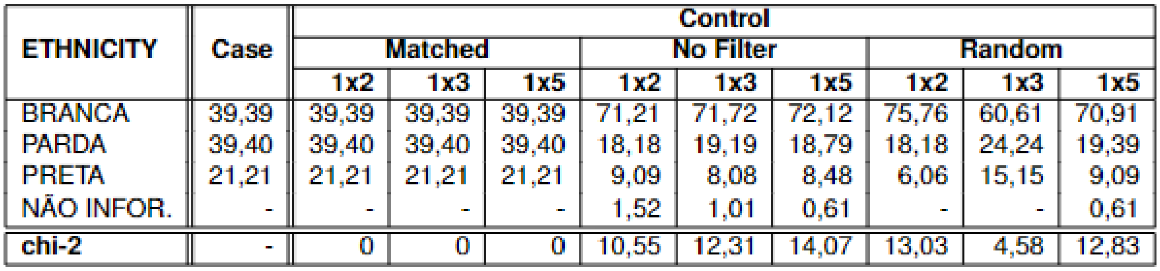
The second attribute in the order of priority was SEX. The behavior of the algorithms was very similar to the behavior for the ETNIA field. The **Matched** algorithm generated perfect (i.e. fully balanced) control files. On the other hand, the frequencies of the files produced by the other two algorithms (i.e. **Random** and **No Filter**) reflect the greater amount of FEMALE values in the complete control file, producing completely unbalanced files.

Finally, the last filter to be applied (i.e. lowest priority) was AGE. The proposed algorithm produced the best results. Although the arithmetic mean was a little higher due to the outliers, the median was only one year higher (i.e. 34 for case base and 35 for control base).

Case 3

The third case study was composed of 33 records in case base and 1277 records in the complete control base. The relevant characteristics (attributes) for this study were the same of the previous one, ETHNICITY, SEX and AGE. The frequencies (in percentage) for each of the ethnicities present in the case base and control bases generated by the 3 algorithms, as well as the chi-square value, can be seen in Figure 5.

Figure 5 – Frequency in percentage (%) for the values of the ETHNICITY attribute (Case 3).



ETHNICITY	Case	Control								
		Matched			No Filter			Random		
		1x2	1x3	1x5	1x2	1x3	1x5	1x2	1x3	1x5
BRANCA	39,39	39,39	39,39	39,39	71,21	71,72	72,12	75,76	60,61	70,91
PARDA	39,40	39,40	39,40	39,40	18,18	19,19	18,79	18,18	24,24	19,39
PRETA	21,21	21,21	21,21	21,21	9,09	8,08	8,48	6,06	15,15	9,09
NÃO INFOR.	-	-	-	-	1,52	1,01	0,61	-	-	0,61
chi-2	-	0	0	0	10,55	12,31	14,07	13,03	4,58	12,83

The performance of the algorithms was very similar to the two previous case studies, where the proposed algorithm achieved a perfect balance and the other two algorithms were not able to produce homogeneous files, as shown by the frequency values. Furthermore, these results are corroborated by the chi-2 values. While the files generated by the **Matched** algorithm reached the ideal value (zero) with a 100% probability of being homogeneous, the chi-2 values for the other two algorithms were very high, with a probability equal to zero of being balanced.

In turn, the behavior of the algorithms for the SEX attribute was very similar to the behavior for the ETNIA field. The **Matched** algorithm generated perfect control files, while the other two algorithms did not achieve a minimally acceptable balance. It is important to highlight that the proposed algorithm managed to perfectly balance the records for the SEX field (second priority) even after balancing the ETHNIC field (first priority).

After choosing the same ETHNICITY and SEX as the case base record, the last filter applied was AGE. The proposed algorithm produced the best results, especially when we compared the median values, which were very close to the case base values. It is worth mentioning that the median is less influenced by outlier values, which can greatly interfere with the arithmetic mean.

Discussion

The proposed algorithm was capable of producing extremely homogeneous control bases, where, in general, the balance decreases according to the priority of the attribute. Furthermore, choosing the order of priority is of fundamental importance to achieve a completely balanced base. For example, in **Case Study 1**, while the **Matched** 2 execution was able to generate a control file with a high degree of balance for all attributes, the **Matched** 1 execution produced a very high imbalance for the STATE field, due to its low priority. Another important observation is that the distribution produced by the **Matched** algorithm slightly worsens as the value of N increases. This behavior is expected since, when increasing the value of N, at each step of the algorithm there will be fewer record options to be used in the next matching, which may result in the algorithm being unable to apply one or more filters. For example, with N=5 after pairing the first record of case base, 5 fewer records will be available to be matched for control base, while with N=2 only 2 records will not be available. This same behavior does not happen in the other two algorithms, as the choice of records does not use any type of priority.

Conclusions

This work proposed and implemented an algorithm to create a homogeneous base for a case-control study (CCS). The algorithm is generic and can be used for any type of case study, where the user chooses the attributes to be paired and their priority order through a GUI. Using this algorithm can greatly help to avoid selection bias, which is one of the main problems with this type of study.

The algorithm was evaluated through 3 real case studies and the results showed the excellent performance of the proposed algorithm, being capable of generating fully balanced control bases. In turn, two other algorithms that choose records randomly or without applying any matching, produced completely unbalanced bases.

Acknowledgements

The authors would like to thank CAPES and CNPq for the support and, also, FAPERJ through the notice APQ1 26/2021 (Process number E-26/211.798/2021).

Referências

1. Tenny, S., Kerndt, C., and Hoffman, M. (2023). Case control studies. StatPearls.

2. Haley, K. E. and Huber, K. E. (2023). Chapter 38 - case-control study. In Eltorai, A. E., Bakal, J. A., Kim, D. W., and Wazer, D. E., editors, Translational Radiation Oncology, Handbook for Designing and Conducting Clinical and Translational Research, pages 223–229. Academic Press

3. Rose, S. and Laan, M. J. (2009). Why match? investigating matched case-control study designs with causal effect estimation. The international journal of biostatistics, 5(1)

4. Pinto, R., Polmann, H., Massignan, C., Stefani, C. M., and de L. Canto, G. (2021). Tipos de vieses em estudos observacionais

5. Dash, S., S. S. S. M. e. a. (2019). Big data in healthcare: management, analysis and future prospects. J Big Data, 6(54).

6. Stanfill, B., Reehl, S., Bramer, L., Nakayasu, E. S., Rich, S. S., Metz, T. O., Rewers, M., Webb-Robertson, B. J., and Group, T. S. (2019). Extending classification algorithms to case-control studies. Biomedical engineering and computational biology, 10

7. Mendes, G. P., Pôrto, L. C. M. S., Lima, C., Santiago, H., Almeida, S., and Sena, A. C. (2023). Análise da prevalência de alelos hla em pacientes com covid-19. Journal of Health Informatics, 15(Especial)

8. Epi Info. Epi Info. Accessed: 2022-04-02