UNIVERSITÀ POLITECNICA DELLE MARCHE

Department of Information Engineering

Master's Degree in Biomedical Engineering

ANALYSIS AND DEVELOPMENT OF A NOVEL ALGORITHM BASED ON MATCHING TECHNIQUE FOR THE ELIMINATION OF CONFOUNDERS IN CASE-CONTROL CLINICAL STUDY.

Advisor:
Prof. Lorenzo Scalise

Candidate:
Ludovica Catani

Co-Advisor:
Prof. Virgilio Paolo Carnielli

Academic Year 2019/2020

INDEX

INTRODUCTION

Research study design is a model of research methods and techniques used to analyse data of specified variable. The majority of research studies evaluate the relation between two different variables to assess if and why one variable is responsible for the changes in the value of the other variable. A clinical study could answer to different research questions, each of which has specific objectives. Different categories of research questions are etiology that assesses the etiological responsibility of the risk factors of a disease, diagnosis that measures the accuracy of diagnostic tests, prognosis that evaluates the natural history of the disease and the power of prognostic factors, and therapy that determines the effectiveness of therapeutic health interventions [1].

There are several types of clinal study design and for the successful execution of the biomedical research is fundamental the appropriate choice of clinical study design. One of the main objectives of clinical study research is to investigate relationship between exposures and healthy populations. Once relationship is identified, the challenge is to determine if and how the exposures truly causes the outcome [2].

This study focuses on epidemiological case-control clinical studies in which the researcher selects the target population and divides it in two groups: one group is composed by subjects that manifest an outcome or disease and it is called case group, the other group is composed by subject of the same target population and so with the same clinical characteristics as case group but with the only difference that they do not present the outcome of interest and it is called comparison or control group. Once created the two groups, the researcher investigates in a backward direction the past exposure to a possible risks factor that could have caused the disease [3].

In case-control clinical study if the study is not conducted carefully and in a proper way the researcher may incorrectly conclude the casual association between two factors exists. This occurs because the observed association is distorted by another factor that was not considered. This distorting is called confounding variable or confounder. In the analysis of case-control studies confounding factors are very common and this is a problem as they falsify the results and influence the effect of the relationship between exposure and outcome creating a spurious associations, for these reasons the researchers try to eliminate or mitigate its effects in clinical studies [4].

The aim of this elaborate is to create a novel algorithm in Matlab environment that eliminates the confounders in a case-control study based on matching technique. Different techniques are used to prevent and manage the effect of confounding variables and nowadays matching technique in case-control studies is performed manually or with proprietary software such as SAS, SPSS and R.

The new algorithm here described is based on a heuristic approach and was tested with different number of sample and compared with another software, SPSS. The algorithm developed is able to match case group with control group and makes the two groups as comparable as possible allowing to minimize differences between confounders. The results obtained were confirmed by applying the algorithm to two different large datasets collect at Neonatal Intensive Care Unit (NICU).

1. Research Study Design: An Overview

For the successful execution of the biomedical research is fundamental the appropriate choice of clinical study design. A clinical study is a scientific process of answering a question using data from a population with a disease to obtain a better understanding of the causes and of the evolution of the disease and to verify if a new preventive, diagnostic or therapeutic clinical approach may be beneficial or effective. Research study design is a model of research methods and techniques used to analyse data of specified variable and the majority of research studies evaluate the relation between two different variables to assess if and why one variable is responsible for the changes in the value of the other variable. Study design can be differentiated according to the direction of inquiry that could be prospective or retrospective. Prospective studies are those in which the researcher looks forward for the outcomes and the participants are involved into the study before they develop the disease. Then they are followed up for certain time to determine whether the outcome of interest occurs [2].

Instead, retrospective studies are those where at the time the study starts outcome of interest has already occurred, so the researcher looks backward and examine exposures to suspected risk factors that probably have caused the outcome. These are known also as case-control studies [5].

There are several types of clinal study design, and they can be subdivided into two main categories: descriptive studies and analytical studies as depicted in figure 1.

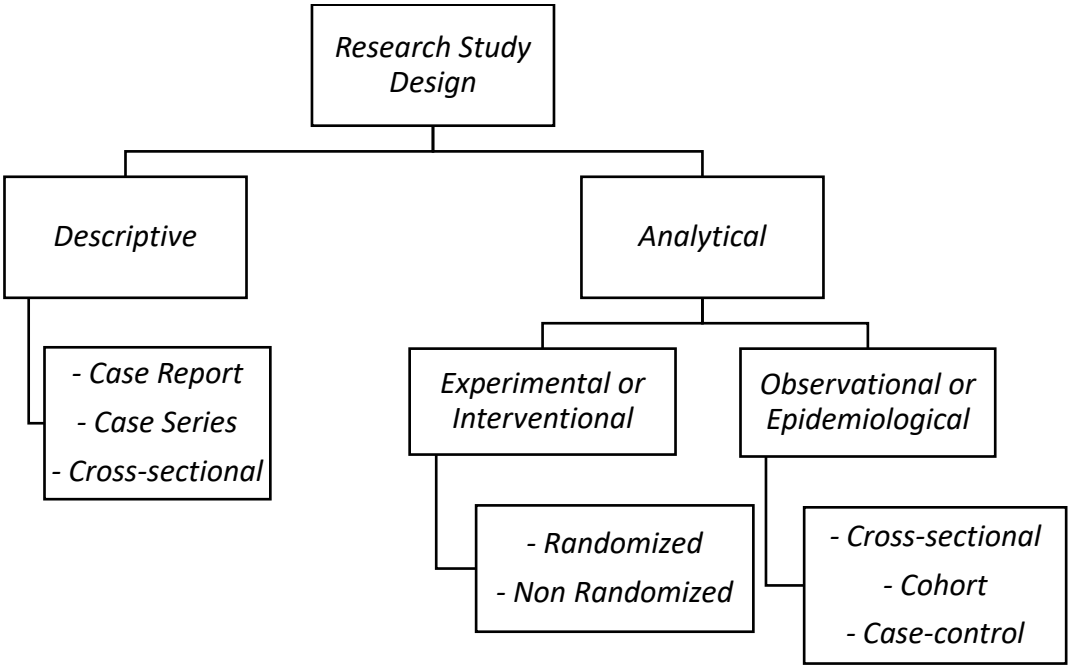

Figure 1. *Research study design hierarchy.*

Descriptive or non-analytical studies are often the first step into a new disease and, as the name suggests, the researcher observes and tries to describe the distribution of variables independently to other hypothesis and without statistical analysis. Normally, descriptive studies include studies in which data are collected on individuals such as the case report that is a detailed report of diagnosis, of treatment or of patient with an uncommon disease and the same exposure. Case series are also a kind of descriptive study, defined as an aggregation of similar case reports with the aim to understand the demographics, prognosis or other characteristics of people who have a particular disease or describe something unusual. The last kind of descriptive study is the cross-sectional that is used to assess the prevalence of a condition in a population and in which the condition of the patient and the potentially related factors are measured at a specific time on a defined population.

Cross-sectional studies can be descriptive when the data are analysed only to determine the distribution of one or more variables, or analytical when the researcher also assesses the relationship between the presence of an exposure and that of an outcome. Cross-sectional studies are inexpensive and easy to conduct compared to other studies, they can provide information on multiple exposures and outcomes and are a good way of assessing the health needs of a populations. However, because the information is collected at a single point in time, it cannot be used to determine whether a particular exposure caused the disease or not [6].

In analytical studies the researcher do not just observe or describe, but tests specific hypotheses, identifies samples of subjects and compares them performing statistical analysis to evaluate information about the effect of an exposure on an outcome, where the exposure is the risk factor while the outcome is the variable that develops as a consequence of the exposure. The two subcategories of analytical studies are experimental studies and observational studies.

The experimental or interventional studies are prospective studies characterized by the active involvement of the researcher that personally decides whether or not a participant will receive the intervention and determines the effect of the exposure in the natural course of events.

The most commonly used experimental studies is the randomized controlled trial (RCT) in which the participants are chosen with the same criteria and then are randomly assigned in two different groups: one experimental group where they receive the intervention that is being studied and the other control or comparison group where they receive an alternative treatment or no treatment [7]. Then, the two group are followed up and compared to see the differences between them in the outcome and theoretically, any difference is

related to the effects of intervention. RCTs are the gold standard of the clinical trial to assess the effectiveness of health intervention [1], they can provide good evidence that the intervention led to an outcome and randomization ensures that both groups have equal chance of receiving the intervention. The disadvantages are that the RCTs are expensive to do and require a large number of participants. In the non-randomized controlled trial there are always experimental group and control groups but differs from RCT because the participants are not randomly assigned to one group, but based on the researcher's convenience or whether a participant can afford a drug or not [8]. Differently from interventional studies, in the observational or also called epidemiological studies the researcher observes the natural effect between risk factor, diagnostic test, treatment or other intervention and the outcome, without trying to change who is or isn't exposed to it [9].

In order to evaluate the diagnostic accuracy of clinical test it is used the cross-sectional study and, as described before in this paragraph, in the analytical cross-sectional studies the researcher only looks at the prevalence of the disease or exposure at one moment in time. It's like a "snapshot" of some individuals in a population diseased or not diseased at one point in time [10]. The purpose is to measure the association between an exposure and a disease or outcome within a defined population.

Cohort and Case-control studies are the two most used types of observational studies that aid in evaluating associations between diseases and exposures [11].

In cohort studies there are different groups of people with shared characteristics and each group have different level of exposure or no exposure. The groups are then followed up over a period of time in order to evaluate the occurrence of an outcome of interest and information about risk factors is

collected. Then the researcher compares the occurrence of an outcome like disease in those who are exposed to a particular risk factor to those who are not exposed to that risk factor. The main measurement used in cohort studies is the relative risk (RR) defined as the ratio between the risk of disease in the exposed group compared to the risk of disease in the unexposed group. RR greater than 1 means that the exposure is associated with an increase of the disease, if RR is equal to 1 means that the risk in the same and if RR is less than 1 indicates that the risk is lower. Typically, cohort studies are prospective studies that means that the outcome hasn't already occurred at the time of experiment and prospective cohort studies are considered the gold standard of observational study design [9]. Sometimes they could be also retrospective studies, the researches may look back at data which have already been collected from participants and the outcome has already occurred [12]. The advantages of cohort studies are that the time sequence of events can be determined, and it is useful when trying to determine what caused the disease. Another advantage is that information about several different outcomes and risk factors can be collected at the same time and this allows for some analysis to be conducted on the data. Disadvantages are the high cost and it can involve a large number of people follower over a long period of time.

Case-control studies are typically retrospective studies in which the researcher chooses two groups of participants with the same clinical characteristics but one group presents the outcome or disease of interest and it is called "case", while the other group is composed by participants without the outcome of interest and it is called "control". The purpose is to investigate, in both groups, past exposure to specific risk factors that are potentially related to the outcome of interest [1]. This kind of study design is commonly used in outbreak

investigations, it is often quick and cheap to do also because it starts with cases it can be used to study uncommon diseases.

One of the challenges in a case-control study is to find suitably matched controls also because in this studies the researcher asks about exposures in the past of the participants and people might not be able to recall their exposures accurately.

In table 1 are summarized all the advantages and disadvantages of the several types of research study design to choose the appropriate model for different biomedical research.

Table 1. *Advantages and disadvantages of research study designs.*

			Advantages	Disadvantages
Analytical	Experimental	RCT	- Reduce bias - Rigorous tool to examine cause-effect relationship	- Expensive - Large number of participants
		Non-RCT	- Larger portion of eligible participants - Greater generalizability	- Subject to manipulations by clinicians
	Observational	Cross-sectional	- Inexpensive - Easy to conduct - Multiple outcomes and exposures can be studied	- No distinguishment whether the exposure precedes or follows the condition.
		Cohort	- Can assess multiple exposures and outcomes - Temporality demonstrated	- Expensive - Time intensive - No good for rare disease
		Case-control	- Quick - Cheap - Good for rare diseases - Look multiple exposures simultaneously	- Prone to bias - Only assess one outcome - Unable to estimate incidence rates
Descriptive		Case report	- Easy to conduct - Ability to make new observations - Educational tool	- Can't prove cause-effect relationship - Can't allow generalizations
		Case series	- Identify rare manifestations of a disease or drug - Provide stronger evidence with multiple cases	- Prone to selection bias - Uncontrolled - Unknown future outcome
		Cross-sectional	- Inexpensive - Easy to conduct - Multiple outcomes and exposures can be studied	- Not suitable for studying rare diseases - Unable to measure incidence

1.1 Case-Control Study

Case-control studies are observational studies because no intervention is attempted, and no attempt is made to alter the course of the disease. They are used to retrospectively determine if there is an association between an exposure and a specific health outcome, so the procedure starts from the outcome or condition or disease to the cause or exposure. The basic design of a case-control study is reported in figure 2.

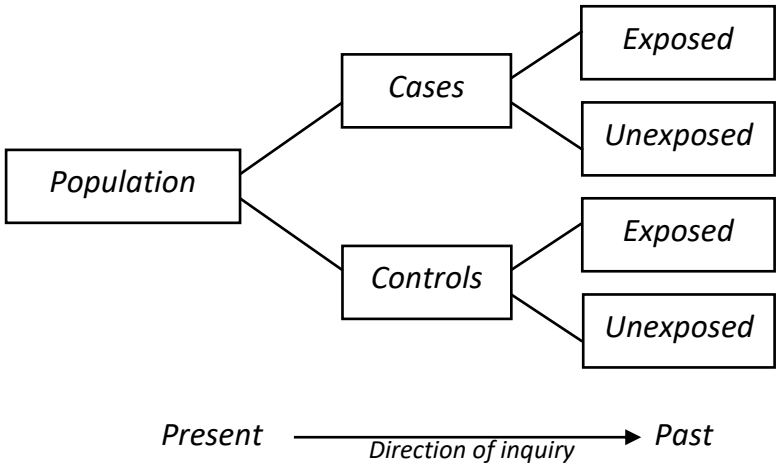

Figure 2. *Design of a case-control study.*

In a case control study, the researcher after randomly selecting a subset of the target population, starts with the identification of cases or people that already have the disease and then creates a comparison group called controls with individuals similar to cases but without the disease. Both groups are interviewed or reviewed of medical records about their previous exposures to different risk factors of interest. After the exposure is determined, the investigator assesses whether exposure is disproportionately distributed between the cases and controls, which may indicate that the exposure is a risk factor for the outcome under study.

In particular, exposure is measured to assess the level of exposure for each individual for the period of time prior to the presentation of the outcome of interest, when the exposure would have acted as a casual factor. As a result, the study is prone to both selection and recall bias. Selection bias may occur when the individuals selected as controls are unrepresentative of the population that represents the cases, and it may also be introduced when the exposed cases are more likely to be selected than unexposed cases. A more serious problem in a case-control study occurs when both case and control participants may not remember the past exposures and, in particular the cases, may remember exposure to a presumed risk factors different from controls. This different recall between cases and controls introduces the so called recall bias, when the medical records of cases are different with respect to controls [13][14][15].

The results of a case-control study can be descripted with a 2×2 table. There are *a* cases who were exposed and *c* cases who were not exposed, consequently there are *b* controls who were exposed and *d* controls who were not, as depicted in table 2 [16][3].

Table 2. *2 × 2 table for case-control study design.*

	CASES	**CONTROLS**
EXPOSED	*a*	*b*
UNEXPOSED	*c*	*d*
TOTALS	*a + c*	*b + d*

In a case-control studies to estimate the strength of the association between exposure and outcome is used the odds ratio or *OR*. It is expressed as the ratio between the odds of being exposed if they were a case and the odds of being exposed if they were a control.

$$OR = \frac{(a/c)}{(b/d)} = \frac{ad}{cb}$$

An odds ratio of more than one means that that risk factor may be related to the development of the disease because cases are more likely to have been exposed to that risk factor than people without the disease (controls). An odds ratio less than one means that the exposure is negatively related with the outcome and an odds ratio equal to one means that there is no association between exposure and disease [14][17].

1.2 Confounder

If the study is not conducted carefully and in a proper way the researcher may incorrectly conclude the casual association between two factors exists. This occurs because the observed associations are distorted by another factor that was not considered and this distorting factor is called confounding variable [18].

One of the most common problem in the design and analysis of observational epidemiological studies and of case-control studies in particular, is the identification of confounding [4]. Confounding arises when the case and control groups are unbalanced with respect to a third factor that influence the outcome of interest and the effect of the third factor gets mixed up with the effect of exposure causing a distortion in the observed association between the outcome and exposure.

Confounders are not measured or controlled as part of the study and a variable is considered a confounder if it meets three conditions such as:

- Confounding variable must be associated with the exposure of interest;
- Confounding variable must be associated with the outcome of interest (i.e. should be a risk factor for the disease);
- Confounding variable could not be a result, or caused by, the exposure of interest.

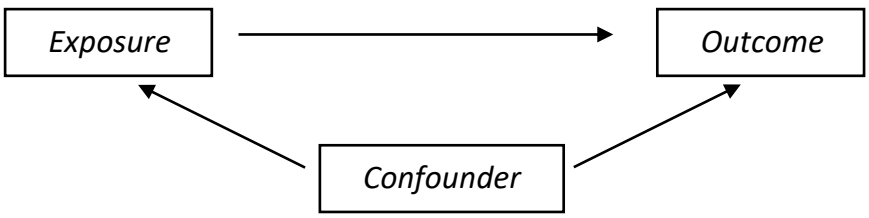

Figure 3. *Relationships between exposure, outcome and a third factor.*

Confounding can be the primary source of bias in case-control studies and can leads to spurious conclusions about a given relationship. Therefore, it is crucial the critical evaluation of the observed association and confounding need to be prevented or mitigated as much as possible in the analysis before looking at the outcome-exposure relationship with the purpose to keep the same effect of all variables both in cases and controls groups [15].

Several statistical techniques could be used to keep the case and control groups as comparable as possible and to increase the validity of the study.

During study design confounding can be prevented with different techniques such as randomization, restriction or matching. Conversely, it can be adjusted after the conclusion of the study with stratification or multi-variate techniques. In case-control study the matching technique is frequently used [19] [20].

1.3 Matching

Matching is an efficient method to address the problem of confounding when the factor suspected to be a confounder increase disease risk but not exposure, and matching efficiency increases with the degree of confounding. The conventional or greedy matching matches control to cases one by one considering random order of cases and when a match is computed, the case and the control are removed from other considerations with the probability to miss some good matches because later cases are not compared to previous matched controls. To obtain an optimal matching in which the sum of dissimilarities between cases and matched controls is a minimum, is needed the evaluation of all possible pairing of cases with controls, but this is an impossible task for any but the smallest studies [21].

Also, of fundamental importance is avoid over-matching that means matching on variables other than those that are a confounder of the disease exposure association or are a risk factors for the disease under study and match only on factors known to be cause of the disease [15].

Matching variable and matching criteria must be set up in advance. Controls can be individually matched that is the most common procedure, or frequency matched. Individual matching searches, for each cases selected for the study, a control with the same specific variable of the cases, this is why it is also called paired matching when one control individually matched to each case or triplet matching when there are two controls individually matched to each case. Frequency matching or group matching select a population of controls such that the proportion of controls with a certain characteristic is equal to the proportion of cases with the same characteristic, so the frequency of a confounder is equal in both groups [3] [22].

1.4 Existing Software

Nowadays most of the time matching technique is performed manually by selecting for each 'exposed' case an 'unexposed' control with the same potential confounder that commonly could be age, sex or socioeconomic status. Also, different software are able to perform the matching in a case-control study and the most commons that perform it automatically are SAS, SPSS and R.

SAS is an analytic software based on a greedy algorithm used to find the closest match and to reduce the computational time. In particular, a greedy algorithm is an algorithmic paradigm used in optimization problems that makes the optimal choice at each step as it attempts to find the overall greedy way to solve the entire problem. It firstly manages the cases one by one randomly without evaluate again the corresponding matched controls for a later case. This method is very fast, and it is perfect for exact matching, but the use of SAS greedy algorithm is not recommended in the multi-object, and this is the main drawback [23].

Another software used for case-control matching is SPSS Statistics by IBM. SPSS is a complete software for data analysis that works with fuzzy machine algorithm to develop optimum solutions for several optimization problems. The last statistical software often used to perform a case-control study is R. With a *MatchIt* function it is able to implements a wide range of matching methods, making possible to reduce the dependence of causal inferences on hard-to-justify, but commonly made, statistical modelling assumptions [24].

2. Materials and Methods

In this study project a heuristic approach based on a genetic algorithm is developed to do matching in case-control study. A heuristic algorithm is a method to solve a problem in faster and more efficient way than traditional methods, it does not guarantee the best solution but is able to provide a good acceptable solution to the problem. One type of heuristic algorithm is the genetic algorithm. A genetic algorithm, also called GA, is an algorithm inspired by the research and the principle of natural selection.

The purpose, in this study, is to offer as similar individual comparison as possible by exploiting the meta-heuristic properties of genetic algorithm.

GA allows to evaluate the starting solutions randomly in an iterative way. At each iteration it selects the solutions based on the loss function and recombine them introducing elements of disorder, so, the selected solutions are able to create new ones in the attempt to converge to optimal solutions. The new elements increase the number of solutions and, as with natural selection, the new more efficient/effective elements (strong elements) replace the less efficient/effective elements (weak elements).

This heuristic technique is usually used to try to solve optimization problems for which no other efficient linear or polynomial complexity algorithms are known. Despite this use, given the intrinsic nature of a genetic algorithm, there is no way of knowing a priori whether it will actually be able to find an acceptable solution to the problem.

The functioning step of GA are summarized in the following table 3.

Table 3. *Functioning step of Genetic Algorithm (GA).*

Step	*Functioning of GA*
1	Random selection of the solution population.
2	Application of the loss function to the solutions belonging to the present population.
3	Selection of the best solutions based on the result of the loss function.
4	Crossing procedure to generate hybrid solutions from the selected solutions.
5	Definition of a new population based on solutions.
6	Iteration of the procedure from point 2 and using the new population created in point 5.
7	Interruption of the procedure.

The functioning step of GA can be applied in a case-control study to find the optimal match. In table 4 are reported the case-control step.

Table 4. *Functioning step of case-control study.*

Step	*Case-control procedure*
1	Random selection of control positions.
2	Calculation of the difference between the confounders of the cases and those of the controls previously selected.
3	Selection of best position sets (absolute mean and rms of smaller differences and/or paired test with larger *p*).
4	Construction of a new dataset with the crossing of the best solutions of point 3.
5	Definition of a new population based on solutions.
6	Iteration of the procedure from point 2 and using the new population created in point 5.
7	Interruption of the procedure.

The loss function lor cost function is the heart of the procedure, it is a function that maps an event onto a real number representing the “cost” associated with the event. It is the one that allows to associate to every solution one or more parameter related to the way in which the solution solves the problem. It is generally associated with the computational performance and thus the temporal performance of the solution.

On these bases, a new algorithm was written and developed in Matlab environment, the new algorithm matches each case with a control selecting the “nearest-neighbor” individual who has not already been selected as a match. It also allows to minimize the difference of confounders and at the same time to have a paired test with *p* as large as possible. In output the loss function favors the datasets that have better results (absolute average and rms of the smallest possible differences and/or test paired with larger *p*).

Was used a multi-objective genetic algorithm (MOGA) because the variable to minimize are multiple and basically, the functioning is the same as a single target GA so it always starts with a creation of a random starting solution of population and with an iterative process it tries to find the optimal solutions but the main difference is that in MOGA there are more than one loss functions to evaluate, so the new algorithm provides multiple results.

2.1 Mathematical Formulation

From a mathematical point of view, there are *i* cases and *j* controls with *n* different possible variables and the aim is to associate one or more control (2) to each case (1).

Cases $\rightarrow F_i^{\,n}$ with *i = 1, …, s* (1)

Controls $\rightarrow G_j^{\,n}$ *with j = 1, …, t* (2)

It is necessary to define the loss function that it is used for the parameters estimation and it is able to evaluate how well the algorithm models the dataset. In this study the loss function was expressed by the mean μ and by the standard deviation σ of the variables as in the formulas (3) and (4), where *z* are the control's indexes that need to be associated with cases.

$$\text{Mean: } \mu_n = \frac{1}{n} \sum_{i=1}^{n} |F_i - G_z| \tag{3}$$

$$\text{Standard Deviation: } \sigma_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (F_i - G_z - \mu_n)^2} \tag{4}$$

The optimal solution is $f(x^*)$ and it is able to minimize the differences between the standard deviations and the means of all *n* variables.

$$f(x^*) \rightarrow \mathbb{Z}^{2n} : (\mu_n, \sigma_n) = f(x^*)$$

$$min(\mu_n, \sigma_n) = f(x)$$

with $x \in X,\ x \geq 1, x \leq t, x \in \mathbb{Z}$.

3. Results

The results were evaluated from three different point of views with a dummy dataset of about 800 samples and two random variables were selected to act as confounders.

First of all, the different heuristic algorithms present in Matlab environment were tested with 450 samples to assess the best heuristic approach. In particular the genetic algorithm was compared with minimax, simulated annealing, and pattern search algorithms.

The minimax algorithm is able to minimize the maximum possible loss, hence the name minimax. In Matlab the *fminimax* function seeks a point that minimizes the maximum of a set of objective functions.

The simulated annealing (*SA* or *Simulatedannealbnd* in Matlab) algorithm is used for optimization problems. An optimization problem is one that has many solutions and each solution has a different score and the goal is to find the best score. *SA* is a metaheuristic approach that tries to find a global minimum when there are different local minima and it is an iterative algorithm in which at each iteration a new point is randomly generated. The distance of the new point from the current point is based on a probability distribution and the algorithm accepts all new points that lower or raise the objective.

The last algorithm tested was the pattern search that, like the other heuristic algorithms, is an optimization method but this try to find the best solution with the lowest error value.

The analysis of variance or ANOVA test was performed with the results for the first confounder, the second confounder and the computing time reported in the following tables and figures.

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	4.36948e+06	3	1456494.416	6319.88	0
Error	7.79056e+06	33804	230.463		
Total	1.216e+07	33807			

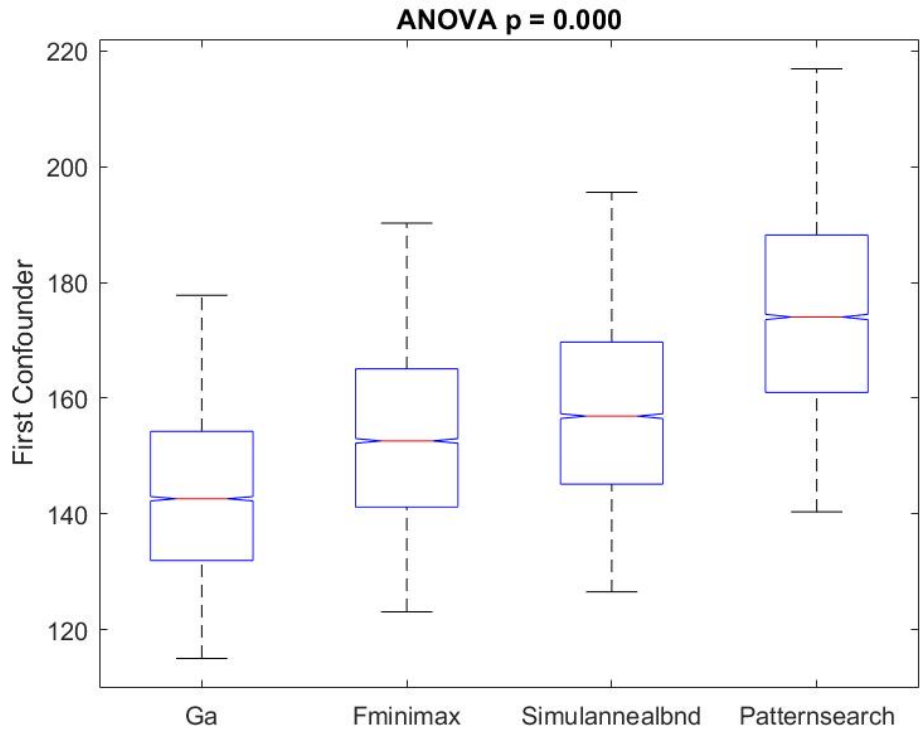

Figure 4. *ANOVA Table and ANOVA Box Plot of the first confounder for each different heuristic approach.*

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	4384.6	3	1461.54	2660.03	0
Error	18573.4	33804	0.55		
Total	22958	33807			

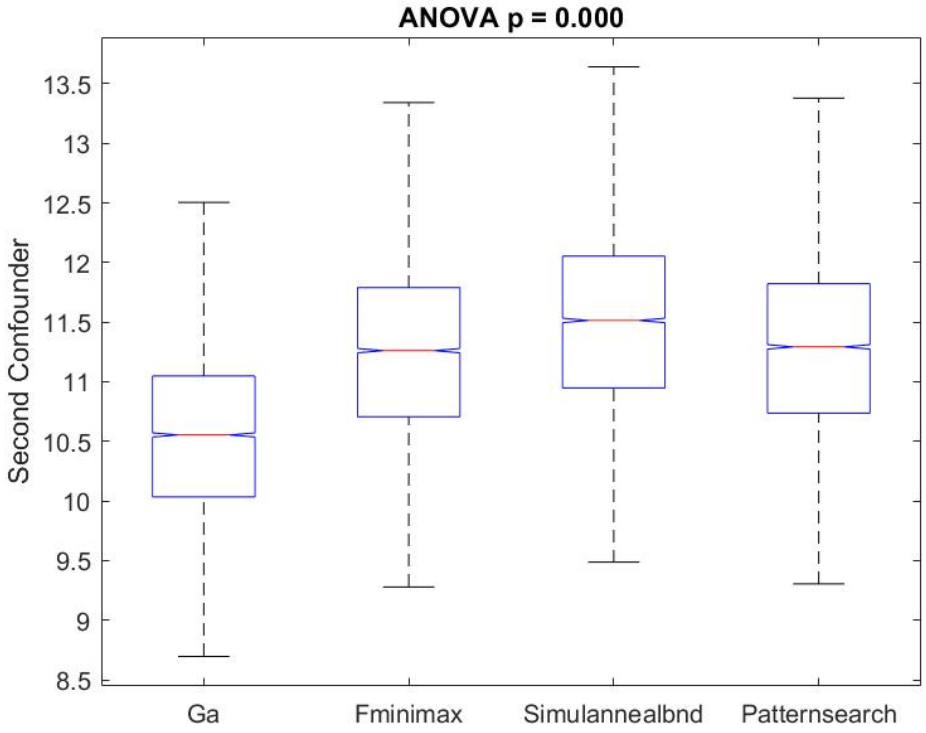

Figure 5. *ANOVA Table and ANOVA Box Plot of the second confounder for each different heuristic approach.*

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	16504	3	5501.34	11.71	4.34201e-07
Error	92111.5	196	469.96		
Total	108615.5	199			

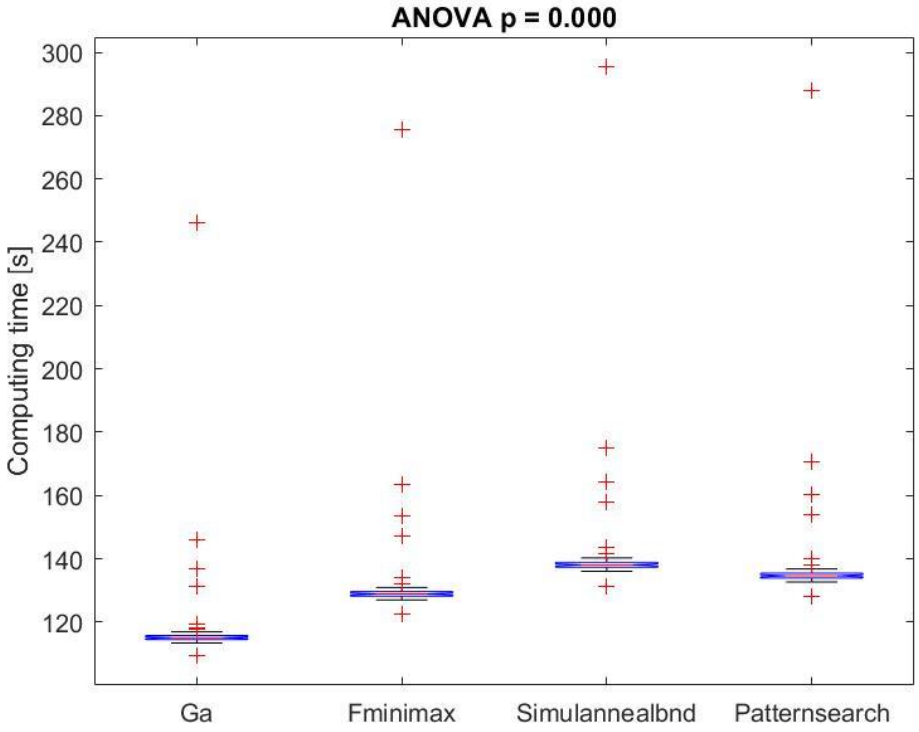

Figure 6. *ANOVA Table and ANOVA Box Plot of the computing time for each different heuristic approach.*

The *p* value that results from the ANOVA tests guarantees that there is a statistically significance difference between the GA and the other heuristic approaches used for comparison.

Table 5. *Mean value for the first confounder, second confounder and execution time.*

	GA	*fminimax*	*Simulatedannealbnd*	*Patternsearch*
Mean first confounder	143.017	153.029	157.319	174.481
Mean second confounder	10.535	11.241	11.494	11.272
Mean time	119.087	133.377	142.904	139.332

In the table above are reported the value for the first and the second confounder and for the execution time for all the heuristic approaches, and finally the GA is the best choice and the more rapid test to perform.

Once GA was chosen as the best heuristic approach, a second trial consists on the evaluation as the number of samples varies. The samples chosen were 50, 200, 250, 400, 450, 500, 650, 800. The variation is represented in the following error bars for the first and the second confounder and for the execution time.

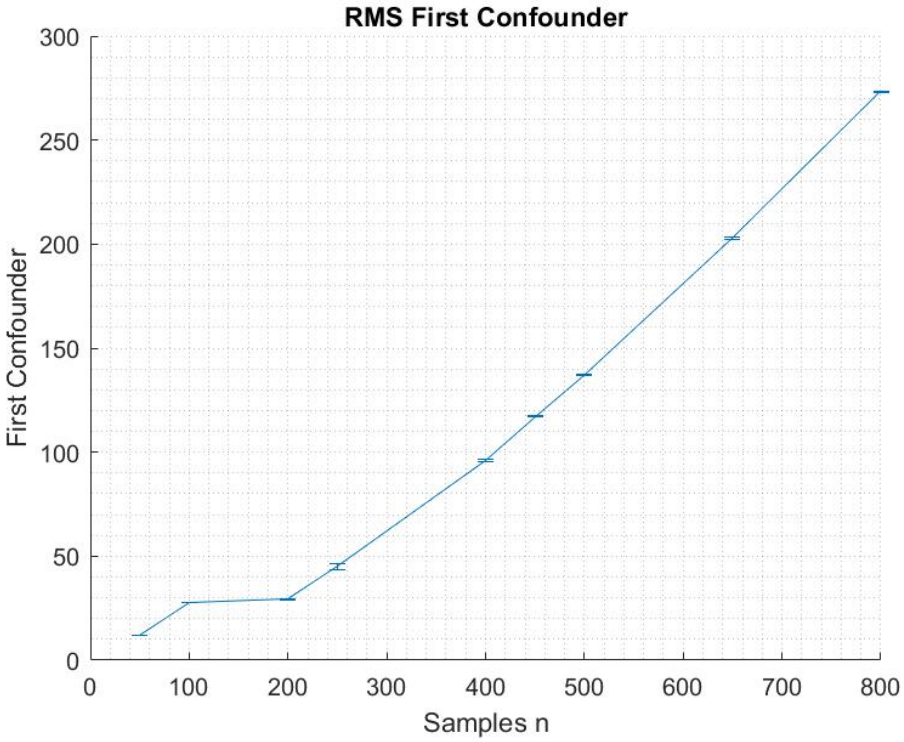

Figure 7. *RMS for the first confounder.*

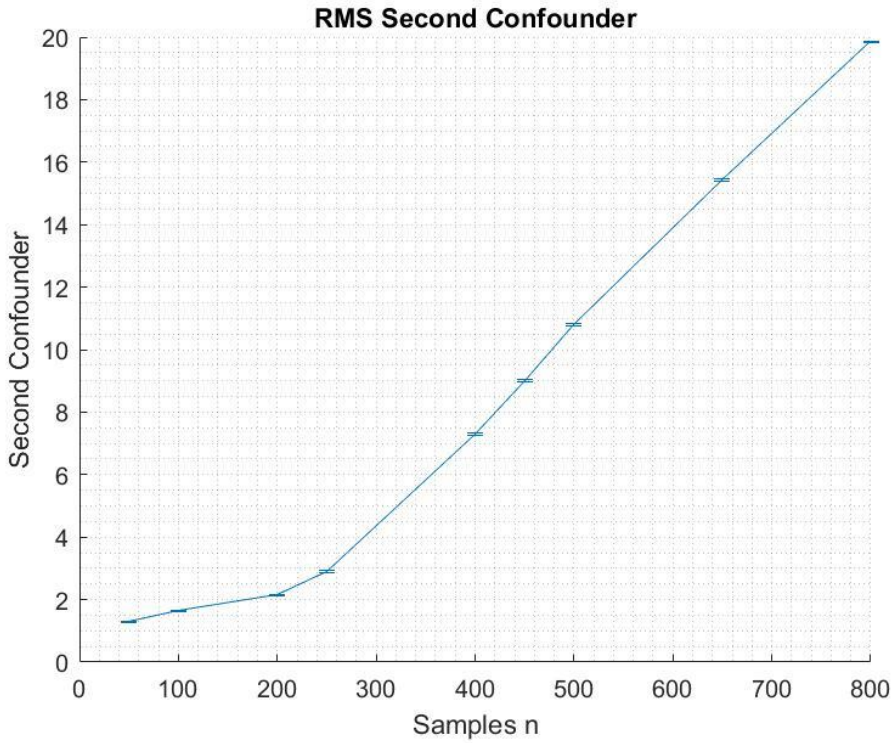

Figure 8. *RMS for the second confounder.*

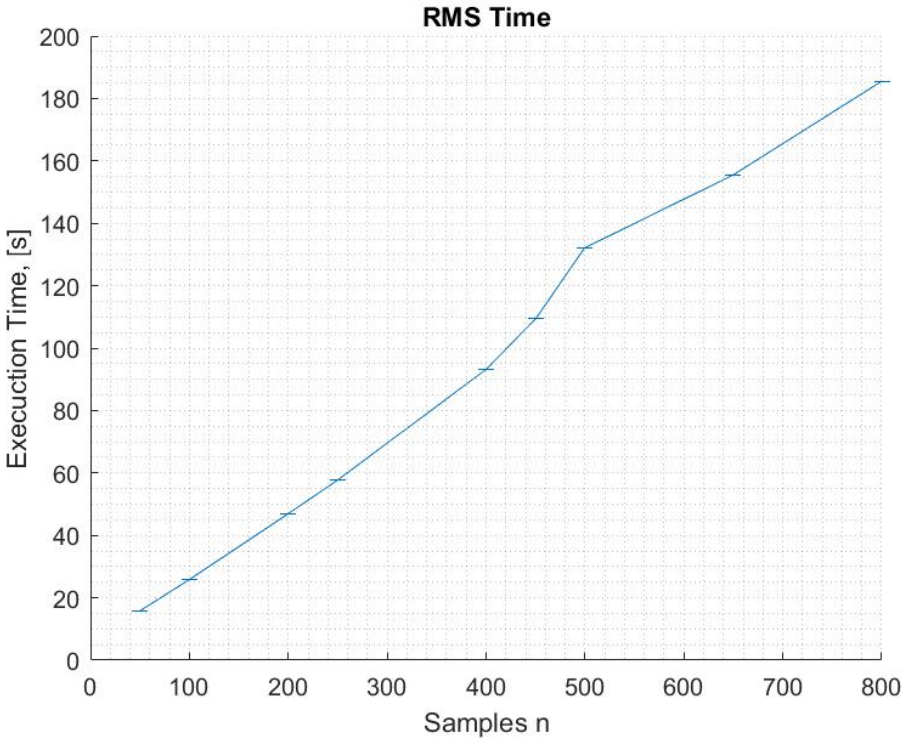

Figure 9. *RMS for the execution time.*

As expected, as the number of samples increases, the variation of the first and second confounder increase, and also the execution time increases due to the increasing number of data to match.

The last evaluation consists in the comparison of the novel heuristic algorithm with SPSS software. A case-control matching and evaluation as the number of samples varies was performed with SPSS. In the following error bars, the SPSS results that oscillates in time are represented with the red line, while the blue line represents the new MOGA.

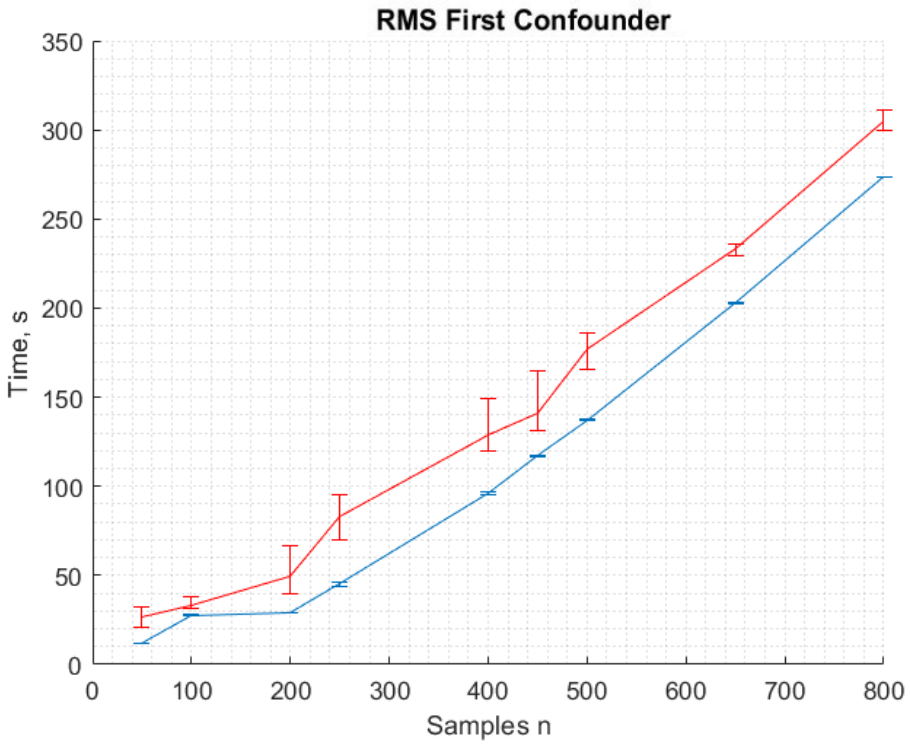

Figure 10. *RMS for the first confounder, SPSS (red line) and new algorithm (blue line).*

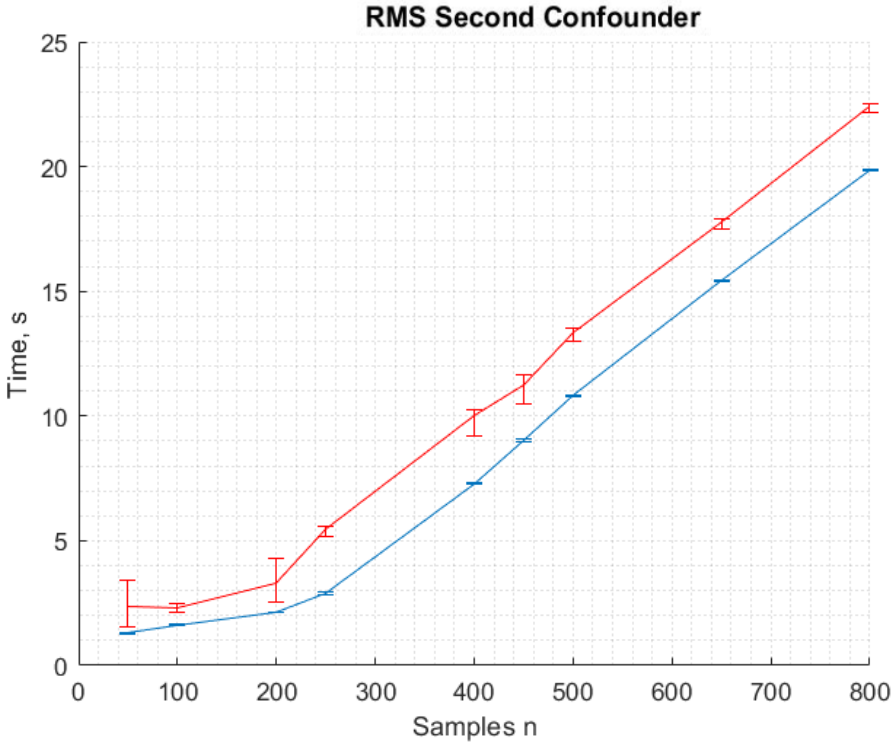

Figure 11. *RMS for the second confounder, SPSS (red line) and new algorithm (blue line).*

The novel algorithm was applied in two different large datasets with 10000 samples each collected at Neonatal Intensive Care Unit (NICU).

The first dataset tested contains the value of partial pressure of carbon dioxide (indicated as PCO_2 and expressed in cmH_2O in the figure) and of foetal haemoglobin (indicated as HfB in the figure). After the application of the novel algorithm the differences between the case group and the control group are reported in the box plot representing the distribution of variables for 1000, 5000, 8000 and 10000 samples.

Figure 12. *Box plot representing the difference between case-control groups with 1000 samples.*

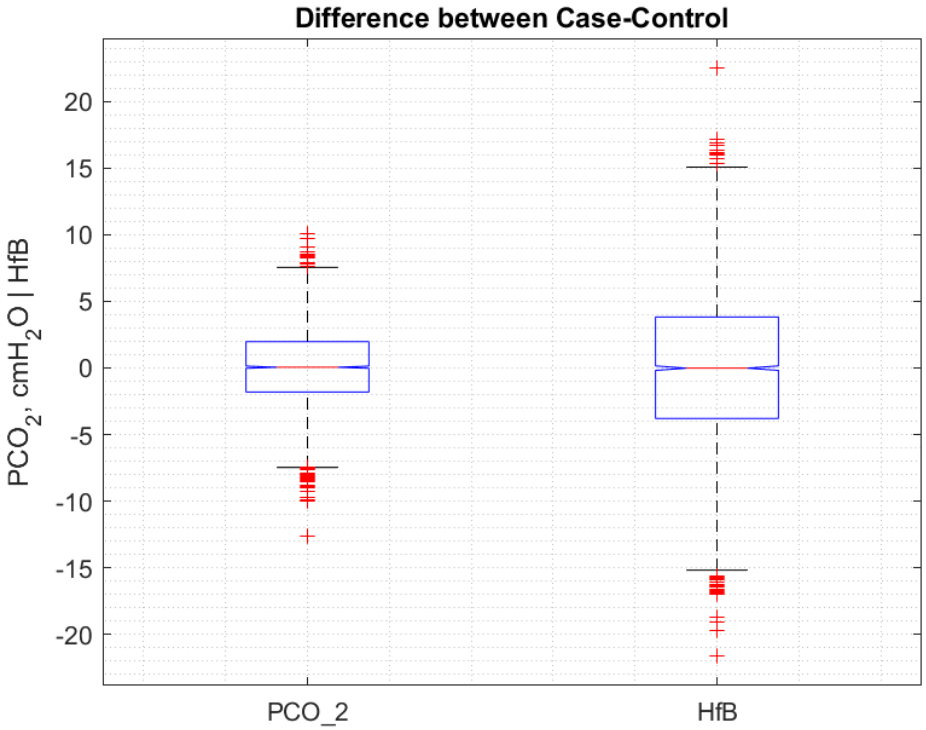

Figure 13. *Box plot representing the difference between case-control groups with 5000 samples.*

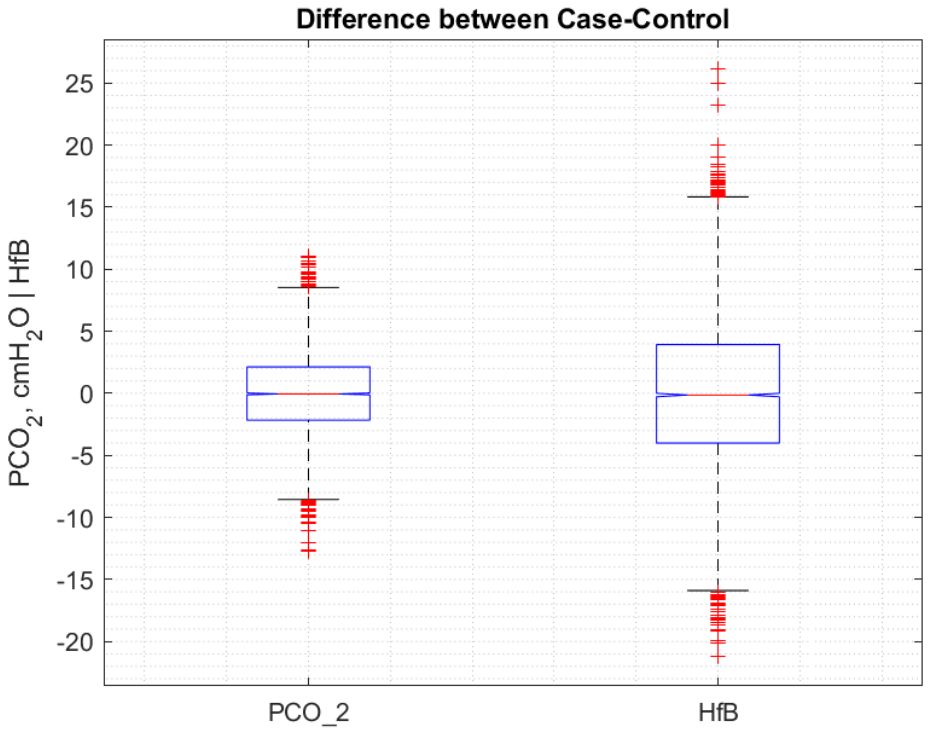

Figure 14. *Box plot representing the difference between case-control groups with 8000 samples.*

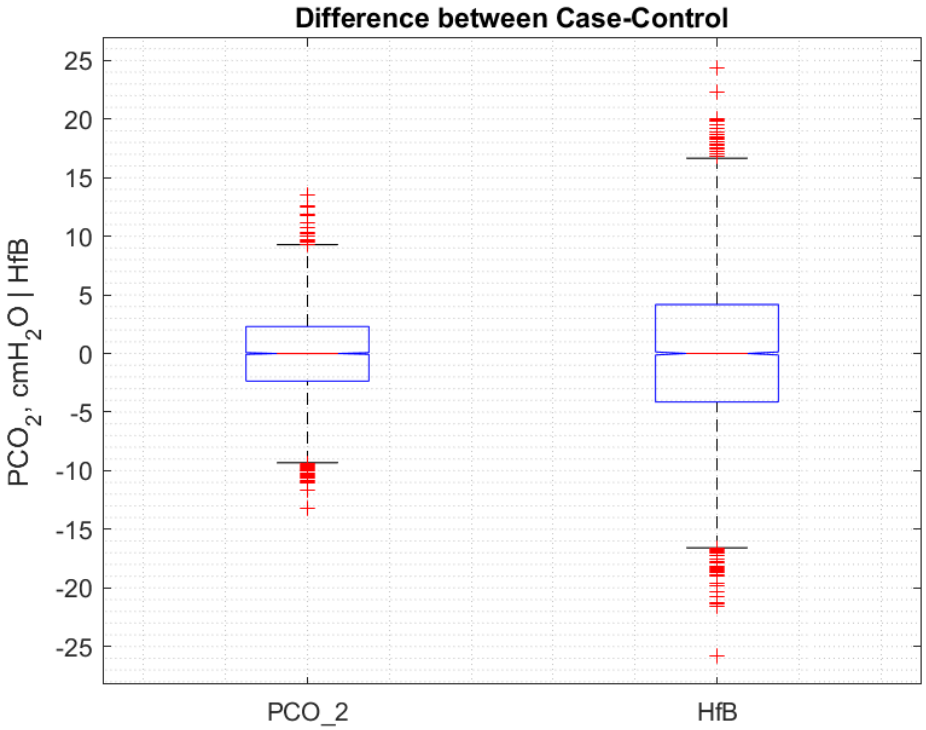

Figure 15. *Box plot representing the difference between case-control groups with 10000 samples.*

In all the trials with different number of samples the mean difference between case group and control group is always around 0, and also the standard deviation is acceptable. The dispersion of values increase as the number of samples increases.

The second dataset includes the oxygen saturation (expressed with SpO_2 in %) with the oxygen partial pressure (indicated as PaO_2 and expressed in cmH_2O). After the application of the novel algorithm the differences between the case group and the control group are reported in the box plot representing the distribution of variables for 1000, 5000, 8000 and 10000 samples.

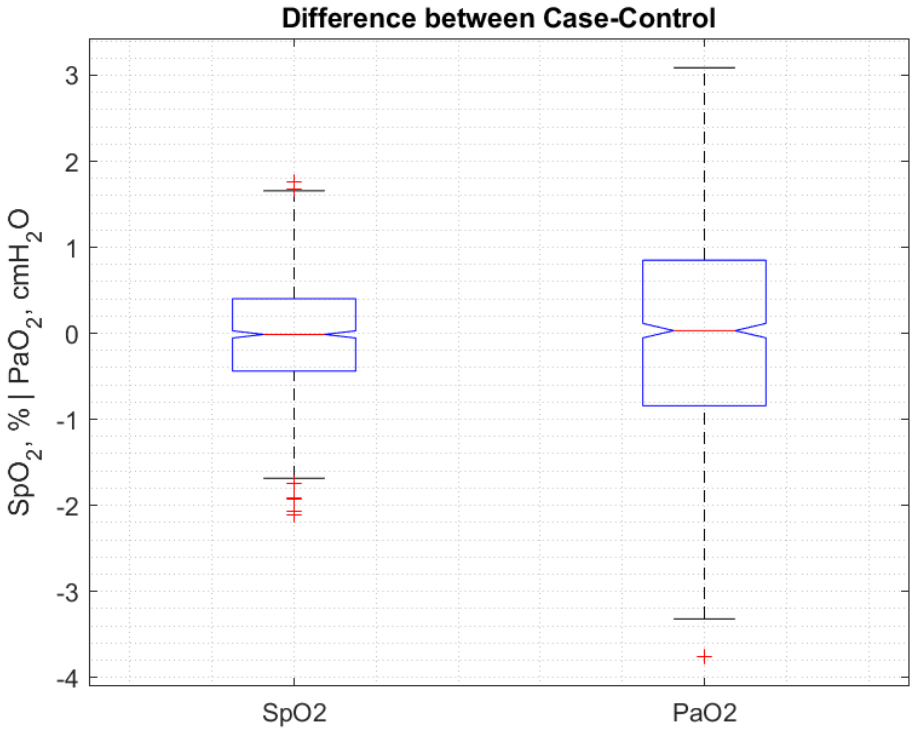

Figure 16. *Box plot representing the difference between case-control groups with 1000 samples.*

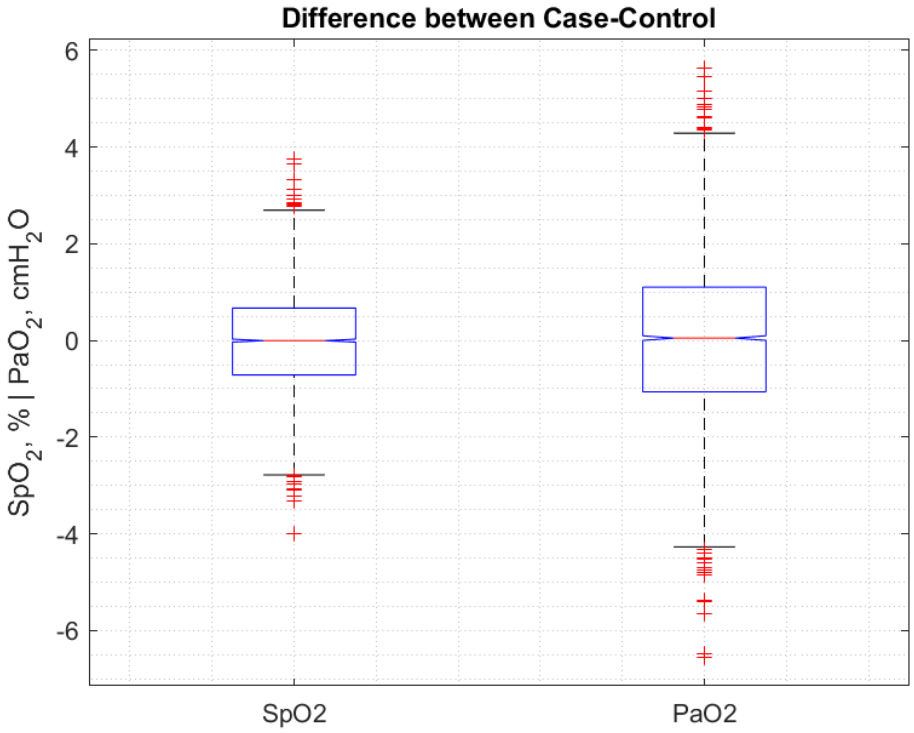

Figure 17. *Box plot representing the difference between case-control groups with 5000 samples.*

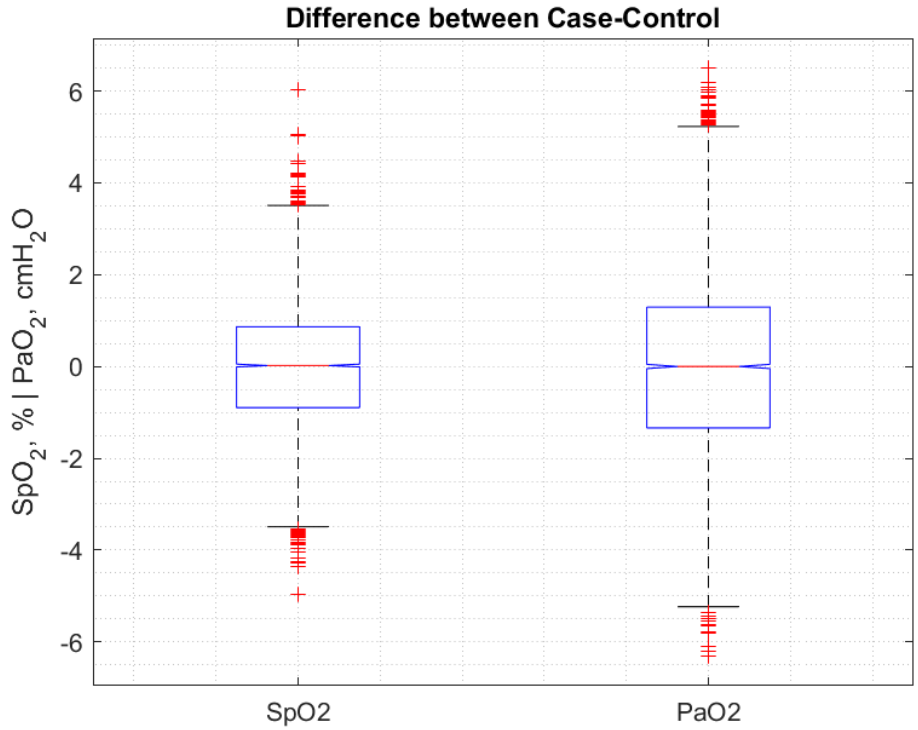

Figure 18. *Box plot representing the difference between case-control groups with 8000 samples.*

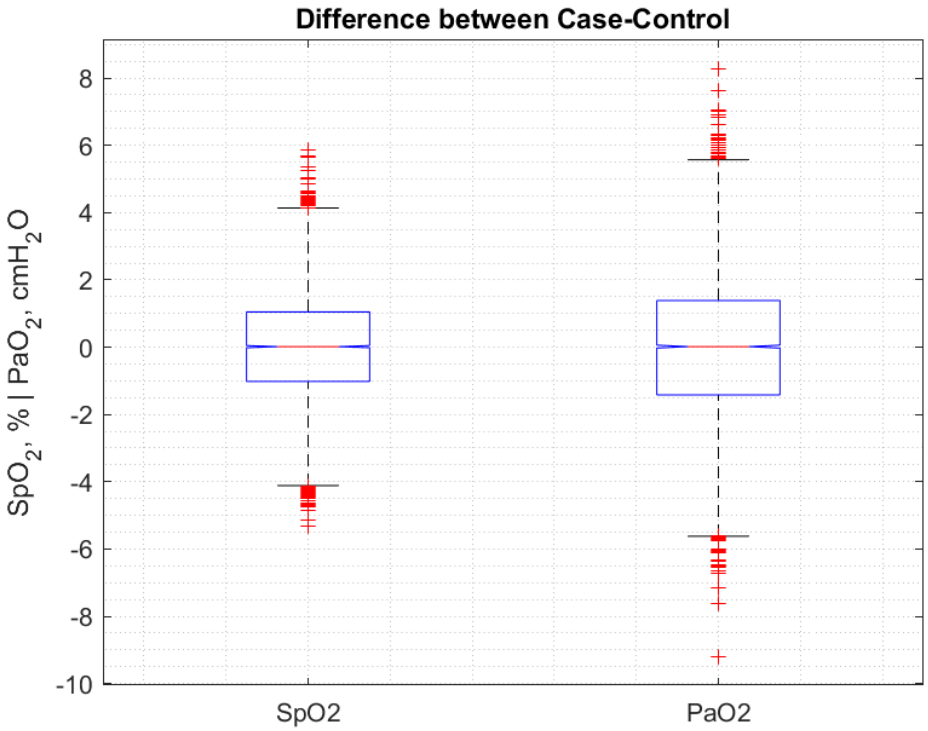

Figure 19. *Box plot representing the difference between case-control groups with 10000 samples.*

Also in this dataset in all the trials performed with different number of samples the mean difference between case group and control group is always around 0, and the standard deviation is acceptable. As expected, the dispersion of values increase as the number of samples increases.

These results are important to state that the novel algorithm based on matching technique makes the two groups as comparable as possible and therefore prevents the influence of confounding variables as much as possible.

4. Conclusion

Confounding can be the primary source of bias in case-control studies and can leads to spurious conclusions about a given relationship. Therefore, it is crucial the critical evaluation of the observed association and confounding need to be prevented or mitigated as much as possible in the analysis before looking at the outcome-exposure relationship with the purpose to keep the same effect of all variables both in cases and controls groups.

The novel algorithm developed in Matlab environment based on the heuristic approach of the genetic algorithm is able to perform matching between case and control groups in a case-control study. The new MOGA was tested with different number of samples and was calculated the RMS for two different variable, chosen randomly to act as confounders, with the results that as the number of samples increases also the RMS increases due to the increasing numbers of case-control pairs to match. The new algorithm is able to match each case with a control selecting the “nearest-neighbor” individual who has not already been selected as a match. It also allows to minimize the difference of confounders and at the same time to have a paired test with *p* as large as possible. In output the loss function favors the datasets that have better results (absolute average and rms of the smallest possible differences and/or test paired with larger *p*). It gives, as output, many optimization solutions allowing the researcher to choose the best one for the study. Finally, it is able to make the two groups as comparable as possible keeping the same effects for all the variables and thus preventing the confounding effects, the results obtained were confirmed by applying the algorithm to two different large datasets collect at Neonatal Intensive Care Unit (NICU).

BIBLIOGRAPHY

[1] N. David, “Pillole Di Metodologia Della Ricerca,” *Gimbe news*, vol. 3, no. I, pp. 15–16, 2010.

[2] P. Ranganathan and R. Aggarwal, “Study designs: Part 1-An overview and classification,” *Perspect. Clin. Res.*, vol. 9, no. 4, pp. 184–186, 2018.

[3] K. A. McBride, F. Ogbo, and A. Page, *Epidemiology, textbook*. 2019.

[4] T. Yanagawa, “Case-control studies: Assessing the effect of a confounding factor,” *Biometrika*, vol. 71, no. 1, pp. 191–194, 1984.

[5] D. R. Hess, “Retrospective studies and chart reviews.,” *Respir. Care*, vol. 49, no. 10, pp. 1171–1174, 2004.

[6] R. Aggarwal and P. Ranganathan, “Study designs: Part 2 - Descriptive studies,” *Perspect. Clin. Res.*, vol. 10, no. 1, pp. 34–36, 2019.

[7] Kendall JM, “Designing a research project: randomised controlled trials and their principles.,” *Emerg. Med. J.*, pp. 164–168, 2003.

[8] P. Ranganathan and R. Aggarwal, “Study designs: Part 5 - Interventional studies (III),” *Perspect. Clin. Res.*, vol. 11, no. 1, pp. 47–50, 2020.

[9] M. S. Thiese, “Observational and interventional study design types; an overview,” *Biochem. Medica*, vol. 24, no. 2, pp. 199–210, 2014.

[10] B. Hernández and H. E. Velasco-Mondragón, “Cross-sectional studies,” *Salud Publica Mex.*, vol. 42, no. 5, pp. 447–455, 2000.

[11] A. Sartori, M. Abdoli, and M. S. Freedman, “Can we predict benign multiple sclerosis? Results of a 20-year long-term follow-up study,” *J. Neurol.*, vol. 264, no. 6, pp. 1068–1075, 2017.

[12] P. Ranganathan and R. Aggarwal, “Study designs: Part 3 - Analytical observational studies,” *Perspect. Clin. Res.*, vol. 10, no. 2, pp. 91–94, 2019.

[13] “International Agency for Research on Cancer Statistical Methods in,” *Cancer*, no. 32, 1980.

[14] K. F. Schulz and D. A. Grimes, “Case-control studies: Research in reverse,” *Lancet*, vol. 359, no. 9304, pp. 431–434, 2002.

[15] N. E. Breslow, “Case-Control Studies.”

[16] E. B. Dupépé, K. P. Kicielinski, A. S. Gordon, and B. C. Walters, “What is a case-control study?,” *Clin. Neurosurg.*, vol. 84, no. 4, pp. 819–826, 2019.

[17] Jeffrey J. Walline, *Designing Clinical Research: an Epidemiologic Approach, 2nd Ed.*, vol. 78, no. 8. 2001.

[18] “Chapter III: General Considerations for the Analysis of Case-Control Studies,” *Statistics (Ber).*, pp. 83–119.

[19] S. Rose and M. J. Van Der Laan, “Why match? Investigating matched case-control study designs with causal effect estimation,” *Int. J. Biostat.*, vol. 5, no. 1, 2009.

[20] K. J. Jager, C. Zoccali, A. MacLeod, and F. W. Dekker, “Confounding: What it is and how to deal with it,” *Kidney Int.*, vol. 73, no. 3, pp. 256–260, 2008.
[21] J. B. Cologne and Y. Shibata, “Optimal case-control matching in practice,” *Epidemiology*, vol. 6, no. 3, pp. 271–275, 1995.
[22] M. A. De Graaf, K. J. Jager, C. Zoccali, and F. W. Dekker, “Matching, an appealing method to avoid confounding?,” *Nephron - Clin. Pract.*, 2011.
[23] B. Snoeijer and E. Heintjes, “Simple and Efficient Matching Algorithms for Case-Control Matching,” *PhUSE*, 2014.
[24] D. E. Ho, K. Imai, G. King, and E. A. Stuart, “Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference,” *Polit. Anal.*, vol. 15, no. 3, pp. 199–236, 2007.