University of Pannonia
Faculty of Information Technology
Doctoral School of Information Science and Technology

SUPPORTING DATA ANALYSIS OF RETROSPECTIVE HEALTH EXAMINATIONS WITH DATA SCIENCE METHODS

DOI:10.18136/PE.2024.885

PhD Dissertation

by

Szabolcs Szekér

Head of Doctoral School
Dr. Ferenc Hartung, PhD

Supervisor
Dr. Ágnes Vathy-Fogarassy, PhD

Veszprém
Hungary
2024

Supporting data analysis of retrospective health examinations with data science methods

Thesis for obtaining a PhD degree in the Doctoral School of Information Science and Technology of the University of Pannonia

in the research field of Computer Sciences

Written by Szabolcs Szekér

Supervisor(s): Dr. Ágnes Vathy-Fogarassy, PhD

propose acceptance (yes / no)

…….
Dr. Ágnes Vathy-Fogarassy, PhD

As reviewer, I propose acceptance of the thesis:

Name of Reviewer: …..... …..... yes / no

…….
signature

Name of Reviewer: …..... …..... yes / no

…….
signature

The PhD-candidate has achieved ….....% at the public discussion.

Veszprém,

…….
Chairman of the Committee

The grade of the PhD Diploma …..... (…….. %)
Veszprém,

…….
Chairman of UDHC

Köszönetnyilvánítás

Ezúton is szeretnék köszönetet mondani mindazoknak, akik hozzájárultak jelen disszertáció elkészüléséhez.

Először is szeretnék köszönetet mondani családomnak. Édesanyámnak, aki lehetővé tette, hogy elkezdhessem akadémiai pályafutásomat. Kislányomnak és feleségemnek az emberi támogatásukért.

Szeretnék köszönetet mondani témavezetőmnek, Dr. Fogarassyné dr. Vathy Ágnesnek, aki megismertetett az egészségügyi informatika rejtelmeivel. Köszönöm szakmai és emberi támogatását, melyeket már az alapszakos képzésem óta kaptam. Továbbá szeretnék köszönetet mondani Dr. Fogarassy Görgynek. Az ő egészségügyi tudása és szakmai segítsége nélkül a dolgozatban bemutatott eredmények nagy része nem születhetett volna meg.

Végül, de nem utolsó sorban, szeretnék köszönetet mondani volt és jelenlegi kollégáimnak, valamint szaktársaimnak és barátaimnak a sok éves emberi és szakmai támogatásért. Külön köszönet Rátosi Márknak, aki már a rögös út legeleje óta mellettem volt.

Abstract

My thesis is about the utilisation of health data assets, in which I focused on two different topics: one being control group selection while the other is extracting information from medical documents.

Observational studies are often based on case-control studies in which the conclusion (for example, the effect of a drug on healing) is drawn based on the comparison of case (treated) and control (untreated) groups. The basic criterion for the proper execution of case-control studies is the selection of appropriate case and control groups. In retrospective studies, the treated group is usually predetermined, and a control group must be selected for the study, in which the individuals are very similar to the subjects of the case group in terms of their basic characteristics that influence the investigated question. In my thesis, I proposed two new nearest-neighbour-based control group selection methods, which perform the selection of individuals in the original n-dimensional feature space. I tested the effectiveness of the proposed methods with Monte Carlo simulations using self-proposed dissimilarity measures and other widely used similarity measures.

In everyday medical practice, the results of echocardiograms are usually recorded in the form of unstructured text, from which extracting relevant information is a challenging task. To support this information extraction, I developed a text mining-based information extraction method that automatically identifies and standardises the descriptions of the heart ultrasound measurement in the findings, and then stores the extracted and standardised measurement descriptions together with the measurement results in a structured form. Through case studies based on large data sets, I have shown that the proposed method can be used to extract measurement results from echocardiography documents with high reliability without performing a direct search or having detailed information about the structure of the document and data recording habits. The proposed methodology effectively handles spelling errors, abbreviations, and varied terminology used in descriptions.

Absztrakt

Dolgozatom az egészségügyi adatvagyon hasznosításáról szól, melyben két különböző témára fókuszáltam: az egyik a kontrollcsoport-kiválasztás, a másik pedig információ kinyerése orvosi dokumentumokból.

A megfigyeléses vizsgálatok gyakorta olyan eset-kontroll vizsgálatokon alapulnak, melyekben a következtetést (például egy gyógyszer hatását a gyógyulásra) az eset (kezelt) és a kontroll (kezeletlen) csoportok összehasonlítása alapján vonják le. Az eset-kontroll vizsgálatok megfelelő végrehajtásának alapvető kritériuma a megfelelő eset- és kontrollcsoportok kiválasztása. A retrospektív vizsgálatok során a kezelt csoport általában előre adott, és a vizsgálathoz olyan kontrollcsoportot kell kialakítani, amelyben az egyedek a vizsgált kérdést befolyásoló alaptulajdonságaik tekintetében nagy mértékben hasonlítanak az esetcsoport alanyaihoz. Dolgozatomban két új legközelebbi szomszéd alapú kontrollcsoport kiválasztási módszert javasoltam, amelyek az egyedek kiválasztását az eredeti n-dimenziós tulajdonságtérben végzik el. A javasolt módszerek hatékonyságát Monte Carlo szimulációkkal teszteltem az általam javasolt különbözőségi mérőszámok és más, széles körben használt hasonlósági mérőszámok felhasználásával.

A mindennapi orvosi gyakorlatban a szívultrahang vizsgálatok eredményeit általában strukturálatlan szöveg formájában rögzítik, amelyekből a releváns információk kinyerése kihívásokkal teli feladat. Ezen információkinyerés támogatására kidolgoztam egy olyan szövegbányászaton alapuló információkinyerési módszert, amely a leletekben automatikusan azonosítja és egységesíti a szívultrahang mérések leírását, majd a kinyert és egységesített mérési leírásokat a mérési eredményekkel együtt strukturált formában tárolja. Nagy adathalmazon alapuló esettanulmányok révén kimutattam, hogy a javasolt módszerrel nagy biztonsággal nyerhetők ki a mérési eredmények az echokardiográfiás dokumentumokból anélkül, hogy közvetlen keresést végeznének, vagy részletes információval rendelkeznénk a dokumentum felépítéséről és az adatrögzítési szokásokról. A javasolt módszertan hatékonyan kezeli a helyesírási hibákat, a rövidítéseket és a leírásokban használt változatos terminológiát.

Abstrakt

In meiner Dissertation geht es um die Nutzung von Gesundheitsdatenbeständen, in der ich mich auf zwei verschiedene Themen konzentriert habe: die Auswahl von Kontrollgruppen und die Extraktion von Informationen aus medizinischen Dokumenten.

Beobachtungsstudien basieren häufig auf Fall-Kontroll-Studien. Bei solchen Studien wird die Schlussfolgerung (z. B. die Wirkung eines Arzneimittels auf die Heilung) auf der Grundlage des Vergleichs der Fallgruppe (behandelt) und der Kontrollgruppe (unbehandelt) gezogen. Das grundlegende Kriterium für die ordnungsgemäße Durchführung von Fall-Kontroll-Studien ist die Auswahl geeigneter Fall- und Kontrollgruppen. Bei retrospektiven Studien ist die behandelte Gruppe in der Regel vorgegeben und es muss für die Studie eine Kontrollgruppe ausgewählt werden, bei der die Individuen den Probanden der Fallgruppe hinsichtlich ihrer grundlegenden Merkmale, die die untersuchte Fragestellung beeinflussen, sehr ähnlich sind. In meiner Dissertation habe ich zwei neue, auf dem nächsten Nachbarn basierende Kontrollgruppenauswahlmethoden vorgeschlagen, die die Auswahl von Individuen im ursprünglichen n-dimensionalen Merkmalsraum durchführen. Ich habe die Wirksamkeit der vorgeschlagenen Methoden mit Monte-Carlo-Simulationen unter Verwendung selbst vorgeschlagener Unähnlichkeitsmaße und anderer weit verbreiteter Ähnlichkeitsmaße getestet.

Im medizinischen Alltag werden die Ergebnisse von Echokardiogrammen meist in Form von unstrukturiertem Text aufgezeichnet, aus dem die Extraktion relevanter Informationen eine anspruchsvolle Aufgabe darstellt. Um diese Informationsextraktion zu unterstützen, habe ich eine Text-Mining-basierte Informationsextraktionsmethode entwickelt, die die Beschreibungen der Herzultraschallmessung in den Befunden automatisch identifiziert, standardisiert und anschließend die extrahierten und standardisierten Messbeschreibungen zusammen mit den Messergebnissen in strukturierter Form speichert. Durch Fallstudien, die auf großen Datensätzen basieren, habe ich gezeigt, dass die vorgeschlagene Methode verwendet werden kann, um Messergebnisse aus echokardiographischen Dokumenten mit hoher Zuverlässigkeit zu extrahieren, ohne eine direkte Suche durchzuführen oder detaillierte Informationen über die Struktur des Dokuments und Datenaufzeichnungsgewohnheiten zu haben. Die vorgeschlagene Methodik behebt effektiv Rechtschreibfehler, Abkürzungen und verschiedene in Beschreibungen verwendete Terminologien.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional LSTM
CNN	Convolutional Neural Network
EF	Ejection fraction
EMR	Electric Medical Record
EPC	Echocardiography Parameter Candidate
FAIR	Facebook AI Research
GoF	Goodness of Fit
HIS	Hospital Information System
JWD	Jaro-Winkler Distance
LCS	Longest Common Subsequence
LD	Levenshtein Distance
LSTM	Long short-term memory
NER	Named-entity recognition
NLP	Natural Language Processing
PMR	Possible Measurement Result
PS	Propensity Score
PSM	Propensity Score Matching
RE	Relation Extraction
RNN	Recurrent neural network
ROC	Receiver Operating Characteristic
WLD	Weighted Levenshtein Distance
WNNEM	Weighted Nearest Neighbours Control Group Selection with Error Minimization
WNNSA	Weighted Nearest Neighbour Control Group Selection with Simulated Annealing

Formulas

Control group selection

X_C	set of potential candidates
X_T	case group
X_{UT}	control group
f_i	i-th descriptive feature
$\mathbf{X}_i$	i-th individual
x_{if}	the value of individual $\mathbf{X}_i$ on variable f
x^*_{if}	the normalised value of the individual $\mathbf{X}_i$ with regard to the f-th covariate
$NN_1(\mathbf{X}_i)$	closest neighbour to individual $\mathbf{X}_i \in X_T$
$NN_2(\mathbf{X}_i)$	the second closest neighbour to individual $\mathbf{X}_i \in X_T$
p_i	the propensity score for the i-th individual
z_i	treatment variable of the i-th individual
p	probability of being exposed
b_i	regression coefficients that describe the relative effects of the covariates (f_i-s) on the status of treatment assignment
w_i	the weighting factor of the i-th covariate
$dissim(A, B)$	the calculated dissimilarity measure of sets A and B
$sim(A, B)$	the calculated similarity measure of sets A and B
NNI^k_{ij}	the Nearest Neighbour Index for $\mathbf{X_i}$ and $\mathbf{X_j}$ individuals in the k-th dimension
GDI^k_{ij}	the Global Dissimilarity Index for $\mathbf{X_i}$ and $\mathbf{X_j}$ elements in the k-th dimension
min_f	the minimum value measured in the f-th dimension taking into account all individuals from $X_T \cup X_C$
max_f	the maximum value measured in the f-th dimension taking into account all individuals from $X_T \cup X_C$

M	set of matched case-control pairs
F_k	the number of possible values along the k-th dimension
$r_{x_{if_k}}$	the ordered rank of the ordinal attribute x_{if_k}
$freq_{kv}^{X_T}$	the absolute frequency of the v-th category in the k-th dimension in the case group
$freq_{kv}^{X_{UT}}$	the absolute frequency of the v-th category in the k-th dimension in the control group
V_k	the set of possible categories along the k-th dimension
$d_{ij}^{(f)}$	the distance of individuals $\mathbf{X}_i \in X_T$ and $\mathbf{X}_j \in X_C$
$E(\mathbf{X}_i)$	Vogel-Korda error function
$p(\mathbf{X}_i, \mathbf{X}_j)$	the probability for selecting the candidate $\mathbf{X}_j \in X_C$ for the individual $\mathbf{X}_i \in X_T$
t	the temperature of the simulated annealing process
e	the energy of the simulated annealing process
M_{i1}	the first element from the i-th pair from M
M_{i2}	the second element from the i-th pair from M
$NN_k(\mathbf{X_i}, Y)$	the k-closest neighbours of $\mathbf{X_i}$ from the set Y
$NN_k(\mathbf{X}_i, X_C)$	the k-size reduced environment for an individual $\mathbf{X}_i \in X_T$
$ActualMatching^{(t)}$	a transient set of matched pairs that the algorithm generates at temperature t
$ActualMatching_i^{(t)}$	an element of $ActualMatching^{(t)}$
$X_{C^*}^k$	the aggregated reduced set of candidates for the k-sized environment
$Dem(\mathbf{X}_j)$	demand set of $\mathbf{X}_j$
$di(\mathbf{X}_j)$	the demand index for $\mathbf{X}_j \in X_{C^*}^k$
$asi(\mathbf{X}_j)$	the alternative selection index for $\mathbf{X}_j$

s_i	string of characters
lev_{s_1,s_2}	the Levenshtein distance of strings s_1 and s_2
$d_w(s_1, s_2)$	the Jaro-Winkler distance of strings s_1 and s_2
$sim_w(s_1, s_2)$	the Winkler similarity of strings s_1 and s_2
$sim_j(s_1, s_2)$	the Jaro similarity of strings s_1 and s_2
l	the length of a maximum 4 characters long common prefix
p	a constant scaling factor with a standard value of 0.1
m	the number of matching characters
t	half of the number of transpositions
$\mathbf{S}_i$	high-dimensional vector representation of string s_i
$sim(\mathbf{S}_1, \mathbf{S}_2)$	the Cosine similarity of $\mathbf{S}_1$ and $\mathbf{S}_2$

Chapter 1

Introduction

The escalating growth of electronic information in the field of healthcare puts a spotlight on retrospective clinical trials. Processing large amounts of data with traditional statistical methods is not always effective.

Due to the unique nature of healthcare and the complexity of the human biological system, for efficient analysis, data mining methods can only be used after area-specific extensions. Improved and healthcare-adapted data science methods can effectively contribute to the analysis of retrospective clinical trials, and they can provide a basis to explore the influencing factors affecting the human biology system. This new knowledge can help physicians achieve individualised medicine.

The aim of my research was to develop such new healthcare-adapted data science methods and algorithms that can effectively contribute to the exploration of information from large, sometimes unstructured healthcare datasets, and extract useful information from them.

My research included the following topics: development of new control group selection methods for retrospective case-control studies; developing new similarity measures for evaluating the results of the control group selection; analysing the effect of missing variables during the control group selection process; and extracting information from large, unstructured healthcare datasets.

1.1 Control group selection

Comparative analysis methods are widely used in observational studies in the field of social sciences [1], natural sciences [2] and engineering [3]. Although these comparison-based scientific analyses significantly differ in the applied methodology

and study design principles [4, 5], due to their comparative nature they have strong scientific evidence [6, 7, 8, 9, 10].

In human comparative cohort studies, people are classified into two independent groups (cohorts) [8], namely into a case group and a control group. The selection of these groups is very important and has a significant impact on the output of the analysis as well. Individuals of these groups have to be similar in many ways (e.g., gender and age distribution), but they have to differ in an examined characteristic property (e.g., patients in the case group are treated with a certain medicine, while individuals in the control group receive placebo) [8, 9, 11].

In prospective study design [5, 11], there are many inclusion and exclusion criteria specified to select the proper individuals into the case and the control groups. Patient-specific data, thought to be important and relevant, are systematically collected and recorded during the whole study period. The main disadvantage of these studies is that the execution of a study sometimes takes up a lot of time.

In contrast, retrospective cohort studies [5, 11] look back in the time and they do not require a long time for collecting data about patients. However, these studies must face the fact that the range of available data is not always complete. The popularity of case-control studies, especially retrospective case-control studies, arises from their relatively inexpensive nature, however, the degree of their evidence is lower than that of randomised trials.

A prerequisite of carrying out appropriate analyses is that the case and control groups have to be similar on covariates (independent variables) that predict group membership (treatment assignment) and affect the examined output. However, fulfilling these requirements is not a trivial task. Many articles have highlighted the importance of the proper implementation of a control group selection method and the effect of unbalanced control groups on the result of analyses [12, 13, 14, 15, 16]. The selection of the case group can be carried out based on the study aims, but the determination of the control group may have difficulties and raises many questions [8, 17]. The reliability of these studies can be improved by (1) increasing the number of cases included in the study, (2) performing thorough data preparation and data cleaning activities, and (3) selecting a proper control group for the case group.

In the literature, various methods have been proposed for selecting a control group. Some of them use different sampling methods (e.g., simple randomised sampling or stratified sampling [18, 19]), while others are based on propensity score matching (PSM) [20, 21]. Nowadays, PSM [20] is the most widely used control group selection method. It is widespread in healthcare analyses [22, 23, 24], and is

gaining ground in social sciences [25, 26, 27] and economics [28, 29, 30, 31].

PSM matches the individuals of the case and control groups based on their propensity score values, which is the probability of the group (treatment) assignment conditional on the observed baseline covariates. Over recent decades, different PSM methods have been proposed (e.g., radius matching, nearest-neighbour matching, stratified matching, kernel matching, Mahalanobis distance matching [32, 21, 33]) to reduce the imbalance of the confounders between the case and control groups [34]. Despite the popularity of these methods, they have also received much criticism [35, 36, 37].

The main limitation of the PSM methods is that they map the feature space into a single value (propensity score), and the matching of the individuals is performed in this compressed space. This can cause the problem of competing risks, which was also highlighted in [38].

Publications [34] and [39] also highlighted that matched-pair analysis has to be performed only when matched individuals are highly correlated, but matching subjects having similar propensity scores does not necessarily result in matched subjects with similar covariate values. Paul Moser in a recently published book wrote that during control group selection, we try to control the influence of the known knowns and the known unknowns [16]. Therefore, if we convert the known things into a compressed 1-dimensional space, which is not able to express as much information as in the original, more informative high-dimensional vector space of the features, the effect of the known covariates cannot be controlled to such a degree.

The selection of covariates is a critical step in case-control studies, and the results of case-control studies rest on a correctly constructed dataset and adequate control group selection. By these considerations, I regard matching in the original n-dimensional vector space or its subspace more suitable than in the 1-dimensional space of propensity scores. The mentioned subspace refers to the covariates which should be included in the propensity score model. Austin [40] and Brookhart [41] recommend that all variables that affect both the exposure of the treatment (group membership) and the outcome of the study should always be taken into account.

For the aforementioned purposes, I developed a novel nearest neighbour-based control group selection method called Weighted Nearest Neighbours Control Group Selection with Error Minimization (WNNEM). The WNNEM method can be seen as a hybrid combination of the PSM method and the nearest neighbour principle, as matching is performed based on the nearest neighbours, but the distances are weighted according to the relevance of the covariates. In [42], I have presented that

the WNNEM method can select more balanced control groups than the greedy PSM method, especially in cases when individuals are characterised only by few covariates and covariates can take only a few values. However, the WNNEM method presented in [42] also has some limitations. On the one hand, it can not handle covariates negatively associated to the treatment assignment; on the other hand, the method can be further improved by utilising probabilistic optimisation for handling more complex problems. Therefore, I proposed a novel nearest neighbour-based control group selection algorithm called Weighted Nearest Neighbour Control Group Selection with Simulated Annealing (WNNSA) [43], which uses simulated annealing for finding the best pairing of the individuals. The proposed algorithm can handle both positive and negative covariates concerning the effect on the probability of the treatment assignment.

The usability of a control group selection method can be defined as how similar the selected control group is to the case group, as the degree of similarity has a significant impact on the evaluation of test results. The evaluation can happen by measuring the similarity of paired individuals from the case and control groups (paired evaluation) or by assessing the similarity of descriptive covariates of the case and control groups (non-paired evaluation). However, the similarity of the covariates of these cohorts is generally not expressed as a single quantitative measure. Only the applied control group selection methods suggest some recommendations on how to perform them in order to be able to select an adequate control group from the available population.

Most of the applied non-paired evaluation methods are Goodness of Fit (GoF) tests (e.g., Kolgomorov-Smirnov test, Bhattacharyya distance, Matusita distance) [44, 45] evaluating the distribution of the two groups. Using a GoF test, it is possible to evaluate a 1-dimensional distribution (that is the similarity of a certain property), but it is nearly impossible for higher dimensions [46].

However, people as the elements of the groups are characterised not by one but by many features. On the other hand, if the elements of the control group are selected by propensity score matching, the similarity of the case and control elements is measured again only in one dimension, namely as the dissimilarities of the propensity scores. As the propensity score is an estimated value, the similarity measurement is made in a lossy compressed 1-dimensional space, and not in the original feature space of the elements.

Contrary to these methods, my aim was to measure the similarity of the case and control groups in the original high-dimensional feature space of the individuals.

For this reason, I proposed three quantitative dissimilarity measures to measure the dissimilarity of the case group and the control group in combination with the previously introduced methods [47, 48]. Two of them evaluate the similarities of case and control groups based on the similarities of the paired individuals and the third one compares the distribution of the characteristic features of the groups. The versatility of the proposed methods was shown on synthetic datasets. Results point out the fact that it is worth considering the proposed measures together to evaluate the similarity of case and control groups and allow researchers to express the degree of similarity of two cohorts quantitatively.

Furthermore, I also analysed the effect of missing dichotomous variables on the deviation of the outcome variable. The analysis was based on a Monte Carlo simulations in which I modelled the effect of omitted variables on the outcome. To measure the bias of the outcome I applied logistic regression-based Propensity Score Matching. I established that in the pessimistic scenario the omitted variables with high significance could greatly affect the value of the outcome variable. This conclusion is based on the revealed linear relationship between the deviation of the outcome and the model accuracy [49]. This analysis drew attention to the important fact that calculations with missing variables can significantly influence the evaluation of case-control studies.

1.2 Information extraction from echocardiography documents

Hospital information systems (HIS) are widely applied information systems for collecting, storing, and managing electronic medical records (EMR). Besides their constantly expanding functionality (e.g., collecting biosensor data), the analysis of information stored in HISs also becomes increasingly important. EMRs are valuable information sources for medical analysis, however they are usually incomplete or redundant, making data mining a difficult and challenging task. The efficiency of information extraction and processing from stored data is significantly influenced by the primary form of data recording. Nowadays, hospital information systems store a large amount of data in a structured form (e.g., personal data, laboratory results), but there are still findings recorded in semi-structured and free-text written format (e.g., anamnesis, echocardiography results). Although the exploitation of information in the data is still typically achieved by human intelligence, artificial

intelligence (AI) algorithms are also gaining ground in this area and help healthcare professionals in solving several domain-specific tasks.

Generally, information extraction from medical texts focuses on the following two tasks: named-entity recognition (NER, or term extraction) and relation extraction (RE). Named-entity recognition refers to the process of identifying particular types of names, terminologies or symbols in documents, while relation extraction identifies the relation between them [50].

Successful term identification is key to getting access to the stored information and the process of identification has been recognised as a bottleneck in text mining. The process of term identification is usually done in three steps: the first step is term recognition; the second step is term classification; and the last step is term mapping [51].

There are two possible approaches to identify terms. The first approach is to directly search for specific terms (e.g., aortic root, ejection fraction) in documents. Direct search can also be extended by pattern search, which requires a priori knowledge about the structure of the processed text (e.g., use of colon between terms and values, order of terms, various expletives). With this extension, it becomes possible to recognise terms and their measured value (e.g., aortic root: 27 mm) together. Other term extraction methods also exist which utilise classical text mining techniques. These text mining-based solutions simply collect every occurrence of word sequences that are possibly valid terms. However, these methods require a text preprocessing phase (including text cleaning), and term candidates must be identified and mapped onto a dictionary after term extraction.

In the literature, several studies have been published which are engaged in echocardiography report processing [52, 53, 54, 55, 56, 57, 58, 59]. Generally, echocardiography reports can be divided into two parts in terms of diagnostic content: in the first semi-structured part diagnostic results are stored in the form of term-value pairs (e.g., interventricular septum: 14 mm) and in the second part results are recorded as free text written in natural language (e.g., mild left ventricular hypertrophy). Processing echocardiography reports is a nontrivial task as the storage of echocardiography examinations varies across different medical institutes.

The methods proposed in the literature are mostly based on the direct search approach, but some of them apply text mining methods as well. In the published studies, typically only the extraction of one specific parameter is the aim, such as ejection fraction (EF). Garvin et al., Kim et al., and Xie et al. all successfully extracted this parameter from free text documents and described practical extraction

techniques [52, 53, 54]. In [55], a natural language-based method was presented which uses a predefined dictionary, expert rules and predefined patterns to extract echocardiography measurements from documents. In this study, a pattern-matching algorithm was created and tested to extract term candidates from a large set of clinical notes. The presented method relies heavily on pattern matching, but it can also identify possible misspellings and synonyms by iterative extraction. Wells et al. also successfully extracted a set of predefined parameters, including wall thicknesses, chamber dimensions or flow velocities [56]. They applied NLP to parse the most frequently measured dimensions and used outlier analysis to filter out unrealistic values. Toepfer et al. developed and evaluated an information extraction component with fine-grained terminology that enabled them to recognise almost all relevant information stated in German transthoracic echocardiography reports at the University Hospital of Würzburg [57]. Jonnalagadda et al. described an information extraction-based approach that automatically converts unstructured text into structured data, which is cross-referenced against eligibility criteria using a rule-based system to determine which patients qualify for a heart failure with preserved ejection fraction (HFpEF) clinical trial [58]. In [59], Renganathan proposed text mining techniques that enable the extraction of unknown knowledge from unstructured documents.

Going beyond the limitations of the proposed methods, I suggested a generally applicable text mining method [60] for extracting numerical test results with their descriptions from free-text-written echocardiography reports. The proposed method abandons regex-based information extraction and employs corpus-independent text mining techniques to extract information from medical texts. It automatically detects expressions containing textual descriptions of the test results and pairs them with their numerical measurement results. The identification of candidate terms is performed by using fuzzy matching utilising the Jaro-Winkler distance to match them to standardised clinical terms. For finding the most suitable text similarity measure, I analysed different distance metrics in, namely Longest Common Subsequence (LCS), Levenshtein distance (LD), weighted Levenshtein distance (WLD), Jaro-Winkler distance, and cosine distance. My experimental results showed that the Jaro-Winkler can discover the most candidate terms at a given threshold.

The suggested similarity-based mapping makes it possible to handle typos, synonyms and abbreviations flexibly; therefore, the efficacy of the information extraction is significantly increased. Additionally, the proposed method can extract multiple information from the documents by a single search, and a repetitive scan is not

needed. The proposed method is mainly recommended for the rapid processing of large volumes of echocardiographic findings, such as to support medical research or to verify patient selection criteria for clinical trials quickly.

The rest of my thesis is organised as follows. Chapter 2 deals with the problems of control group selection, including novel selection and evaluation methods, while Chapter 3 introduces the proposed text mining method to extract information from echocardiography documents.

Chapter 2

Control group selection

Matching-based control group selection methods aim to select and pair individuals from a set of potential candidates (X_C) to individuals of the case (treated) group (X_T). Individuals $\mathbf{X}_i \in \{X_C \cup X_T\}$ are characterised by n ($n \in \mathbb{N}$) descriptive features (e.g., age, gender, diagnoses) denoted as $f_1, f_2, \ldots, f_n$. Therefore, each subject is denoted as an $\mathbf{X}_i = [x_{if_1}, x_{if_2}, \ldots, x_{if_n}]$ vector of variables, where $i = 1, 2, \ldots, l$ and $l = |X_C \cup X_T|$.

The aim of control group selection methods is to select such an X_{UT} control (untreated) group that is balanced to the case group, meaning that the distributions of the variables (f_i) in both sets are similar. Naturally, X_T and X_{UT} must be disjoint sets, that is, $X_T \cap X_{UT} = \emptyset$. To ensure this requirement, X_T and X_C must also be disjoint ($X_T \cap X_C = \emptyset$).

During my research on control group selection, I focused on three different questions: (i) how to quantify the quality of a control group, (ii) how to select an adequate control group, and (iii) what happens when not all important variables are used when we select a control group.

Chapter 2 is organised as follows. Section 2.1 presents the most widely used control group selection methods and evaluation measures. In Section 2.2, three different dissimilarity measures are presented to answer question (i), Section 2.3 aims to answer the question (ii) while introduces two novel control group selection methods, and question (iii) is discussed in Section 2.4.

2.1 Theoretical background

2.1.1 Control group selection methods

As I mentioned in Section 1.1, various methods exist for selecting a control group, but the most widely applied method is Propensity Score Matching (PSM). In this section, I give a detailed introduction of PSM and summarise the other methods.

2.1.1.1 Propensity Score Matching

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

$$p_i = Pr(z_i = 1|\mathbf{X}_i), \tag{2.1}$$

where p_i denotes the propensity score for the i-th individual and $z_i \in \{0, 1\}$ denotes the treatment variable in such a way that $z_i = 0$ refers to the untreated (control) group and $z_i = 1$ refers to the treated (case) group. Subjects characterised by the same properties have the same propensity scores.

In retrospective observational studies, the true propensity score is unknown and has to be estimated from available data. Usually, it is estimated using a logistic regression model, but other methods have also been examined and used (e.g., recursive partitioning [61], random forests [62], bagging and boosting [63, 64] and neural networks [40, 65]).

When the dependent variable is dichotomous, logistic regression is the most commonly used method to estimate the propensity scores. In this case, treatment status is regressed on the observed baseline covariates and propensity scores are estimated by the fitted model. The multiple linear regression function estimated by the logistic regression model can be calculated as

$$logit(p) = b_0 + b_1 f_1 + b_2 f_2 + \cdots + b_n f_n, \tag{2.2}$$

where

$$logit(p) = ln\left(\frac{p}{1-p}\right) \tag{2.3}$$

and p is the probability of being exposed, furthermore, b_i-s ($i = 1, 2, ..., n$) are the regression coefficients that describe the relative effects of the covariates (f_i-s)

on the status of treatment assignment. The propensity score estimated by logistic regression is calculated as

$$p = \frac{e^{(b_0+b_1f_1+b_2f_2+\cdots+b_nf_n)}}{1+e^{(b_0+b_1f_1+b_2f_2+\cdots+b_nf_n)}}. \tag{2.4}$$

Various propensity score-based matching methods exist that may differ in terms of selection methodology, the ratio of the treated and untreated individuals or the nature of the selection process [66, 67, 68, 69, 70].

Firstly, individuals can be selected into the control group with or without the replacement of the candidates. A general tendency is to apply PSM with replacement when the population from which the control group is selected is too small. Otherwise, matching without replacement is used.

Secondly, the ratio of the number of individuals in the case and control groups can also be varied. One-to-one (1:1) matching is common practice, but in case of large datasets, other implementations, such as one-to-many (1:N) matching, can also be used.

Lastly, the variety of the PS-based matching methods also increases by the fact that during the selection of the individuals, greedy or optimal matching can be applied. In the first case, untreated subject whose propensity score is the closest to the score of a given treated subject is selected and matched. When optimal matching is used, the aim is to minimise the total within-pair difference of the propensity scores, and the pairing is optimised globally [71, 72].

It was shown in a recent article [73], that out of 1000 articles using PSM (published between 1983 and 2015), only $6\,\%$ used any iterative balance checking procedure. In the remaining $94\,\%$ of the articles, simple 1:1 greedy PSM was applied without any balance checking. Therefore, the most widely used implementation of propensity score matching is 1:1 greedy matching without replacement. In the case of 1:1 greedy matching, exactly one subject is paired to each individual of the case group. Furthermore, in this case, matching is based on the nearest neighbour, that is an individual selected for pairing whose propensity score is the most similar to the propensity score of the current individual of the case group. If multiple subjects from the candidates have equally close propensity scores to the propensity score of the sample subject, one of those is selected at random.

As the greedy method does not contain any restrictions concerning the maximum acceptable difference between the propensity scores of the two matched subjects, practical implementations often take into account a threshold parameter for

the selection [74]. Individuals within a certain distance of the propensity scores (caliper size) are matched together, and subjects that fall outside this caliper are neglected. Various suggestions have been made for optimal caliper size in the literature [74, 75, 76], but usually 0.2 of the standard deviation of the logit of the propensity scores is recommended [74].

Despite the popularity of the widely applied 1:1 greedy PSM, it has also got many criticisms [13, 16, 35, 37, 39, 73, 77]. All these articles pointed out that the PSM method in some cases and studies may result in a not well-balanced control group. For example, in [73] the authors highlighted that propensity score matching might increase imbalance [1] even relative to the original data.

Most of the critical comments point to the possible imbalance between the case group and the control group. For example, King and Nielsen highlighted that PSM is blind to the often large imbalance that can be eliminated by approximating full blocking with other matching methods [73]. Moreover, they pointed out that propensity score matching may increase imbalance even relative to the original data. In [78], the authors showed based on four cardiovascular studies that propensity score methods are not necessarily superior to conventional covariate adjustment. Peter C. Austin, who has researched and applied PSM in several fields, also published a review which summarizes the critical appraisals of propensity score matching methods in the medical literature between 1996 and 2003 [77]. He reviewed 47 articles in which propensity score matching was applied, and found two articles that report imbalance on the baseline covariates between the case group and the control group despite proper application of the PSM method.

The main problem with the most widely applied form of PSM presumably originates from the application of dimension reduction on the original feature space: pairing is performed in the 1-dimensional space of the propensity score values, which reduced space might hide the distributions of the original dimensions (features). Covariates that equally affect the probability of treatment assignment (meaning that individuals are assigned to the treated group) also affect the value of the propensity score to the same extent. However, the distribution of these variables may be different, and this difference will no longer appear in the 1-dimensional probability values. As long as the matching is performed in the 1-dimensional space of the propensity scores, these differences can not be taken into account during the matching procedure.

[1] A classification data set with skewed class proportions is called imbalanced.

2.1.1.2 Other methods

In addition to PSM, other control group selection methods are also available, but these methods are not as popular. Such a method is for example the Stratified Matching (SM). SM distributes the individuals into smaller spaces, called strata, and selects pairs stratum-by-stratum. The condition for the successful application of SM is that there should be enough candidates in each stratum to perform the pairing. The disadvantage of this method arises from the difficulty in handling continuous features, as the binning method significantly influences the matching results.

Nearest Neighbour Matching (NNM) is also a simple but effective method. NNM selects pairs based on the Euclidean distance between the elements. However, when the Mahalanobis metric is used instead of the Euclidean distance for distance calculation, we can talk about the Mahalanobis Distance Matching (MDM). Mahalanobis distance is kind of like a scale-free Euclidean distance, and it is very useful in cases when the control group selection has to be done when the individuals are characterised by nominal or ordinal features.

2.1.2 Evaluating the similarity of case and control groups

Evaluating the result of a control group selection method is done by measuring the similarity of the case group and the control group. The similarity of two multidimensional groups is not clearly defined in multidimensional data analysis and therefore, the measurement of the similarity of the case group and the control group is not a trivial task. The classical hierarchical clustering similarity measures (single linkage, complete linkage, average linkage) [79] reflect the similarity of two sets of elements, however, they cannot be used to express the similarity or dissimilarity rate of the groups of cohort-based studies. On one hand, they only consider the similarity of one element pair (single linkage, complete linkage), on the other hand, they evaluate the similarity of the groups by considering the similarity between all possible element pairs (average linkage method). Since the basis for the comparative analysis is that the two study groups have a similar exposure to independent variables, the measurement of their similarity must also be performed some other way.

Consider the following example. There is a case group and a control group, and the distribution of sexes (male or female) is $50-50\,\%$ in each. Can the two groups be considered similar when every male but none of the females in the case group smokes but in the control group the smoking habits are exactly the opposite? Of

course not. In this case, the distribution of smoking habit is the same, but the paired elements differ from each other. However, the opposite, where the paired elements are very similar, but the distributions of the variables are different in the case group and the control group (e.g., individuals in the control group are from 1 to 5 years older than their pairs in the case group) may also happen. As it can be seen, we can only say that the two groups are similar if the distributions of the predictive variables are similar in the two groups, and furthermore, paired individuals are also similar to each other.

Therefore, the similarity of two cohorts must be tested generally, without any pairing of the elements (*distribution-based evaluation*) and the similarity must also be tested by utilising some kind of a pairing method (*pairing-based evaluation*). In the following, the measurement of these basically different similarity aspects are detailed.

In the case of *distribution-based evaluation*, the dispersion of the values of the features is examined in each dimension, and the distributions of the same dimensions are compared. The similarity of the two cohorts can be calculated from the similarities of the distributions of the dimensions. For nominal data, the chi-squared test [80], for ordinal data, the Mann–Whitney U test [81, 82] can be applied. For continuous variables the standardized mean difference (SMD) [83] of the variables can be tested or Goodness of Fit tests can be used. In the case of a normal distribution, the t-test [84], for general cases, the Kolmogorov–Smirnov [85, 86] test can be calculated. The Student's t-test is able to compare the means of two independent samples, but this test assumes that data is normally distributed and the two populations have the same variance. Furthermore, this test compares only the mean values of the datasets, and datasets have to arise from 1-dimensional populations. The Kolgomorov-Smirnov test is also able to compare two probability distributions. Although it is suitable to test the equality of any type of continuous distributions, its application is also restricted to the 1-dimensional vector space. The main drawback of these tests is that they evaluate the similarity of the case group and the control group on only a single covariate. However, people as the elements of the case group and the control group are characterised not by one but many features.

In the literature, there have been some solutions proposed for multidimensional problems as well, but they also have many limitations. For example, the Hotelling T^2 test [87] is the multivariate extension of the two-sample Student's test. The main limitation of this test is that it assumes that each population follows the multivariate normal distribution, making it incapable of handling other types of continuous dis-

tributions and discrete variables. There are similar problems with the recently published methods as well which aim to test the mean vectors of two high-dimensional datasets [88, 89, 90]. In biomedical studies, the Hansen and Bowers test [91] is applied for more complex evaluations. This measure allows the evaluation of the imbalance of all covariates simultaneously.

In case of *pairing-based evaluation*, the elements of the cohorts are paired (a pair contains one individual from the case group and one individual from the control group) and the similarity of the paired elements is evaluated one by one. The similarity of the two groups can be calculated as an aggregated value of the pairwise similarities (i.e. the average or the weighted average of the pairwise similarities). If the size (number of individuals) of the case group and the control group is different, one-to-many assignment can also be used.

The utilisation of the cross-match test [92] is a very interesting approach. This method does not consider the distributions and is based on the adjacency of the individuals. The adjacency is determined as a non-bipartite matching and the similarity of the two groups is expressed as the number of pairs containing one observation from the first distribution and one from the second. This interesting idea is really able to decide whether the two groups of individuals are similar to each other, but in the case of similar groups, the differences of the paired individuals are not expressed. However, since pairing methods are always approximation-based, it is important to express the similarity of the paired individuals quantitatively as well.

Unfortunately, little attention is paid to the evaluation of the goodness of the control group in many studies. However, in the absence of this or in the case of bias in the control group, the results of the comparative analysis are questionable.

For these reasons, three similarity measures are proposed to fill the deficiencies described above. Two of those pair elements of the case group to elements of the control group and evaluate the similarity of these paired elements in different ways, while the third measure is a distribution-based measure which is able to evaluate the similarity of any kind of multidimensional dataset. I believe that by evaluating these measures simultaneously, data analysts can gain a detailed picture of the similarity of the groups partaking in the study and of the nature of the differences between them.

2.2 Novel dissimilarity measures

In this section, the proposed dissimilarity measures are presented and evaluated. Section 2.2.1 introduces the measures while Section 2.2.2 presents the evaluation results.

My aim was to express the dissimilarity of the two cohorts with dissimilarity measures in the range $[0, 1]$. In this way, the similarity of the two groups can be calculated as $sim(X_T, X_{UT}) = 1 - dissim(X_T, X_{UT})$, where $dissim(X_T, X_{UT})$ yields the calculated dissimilarity measure of the two groups.

2.2.1 Definition of the measures

2.2.1.1 Measures for pairing-based evaluation

Nearest Neighbour Index

The first measure is called Nearest Neighbour Index (NNI) and it is quite strict. NNI checks for each attribute whether the case-control entity pairs are the closest neighbours to each other on that attribute. However, the index does not measure the distance from the closest value along the given dimension. As element-pairs can be determined by any kind of matching method, the index is applicable for any kind of case-control pairing-based assignment. NNI is calculated the following way.

- For continuous features, the dissimilarity is 0 if and only if the sample-control pair is the closest to each other pursuant to the examined attribute, otherwise it is 1.
- For categorical and ordinal features the dissimilarity is 0 if the values of the attributes of the individuals are identical, otherwise 1. In case of categorical features there is no order between the possible values, so it is acceptable. In case of ordinal features we lose information about the magnitude of difference between the two values. As the name suggests, Nearest Neighbour Index only considers perfect correspondence, and the difference between the values is not measured.

The Nearest Neighbour Index can be formally described as:

Dissimilarity for continuous features:

$$NNI_{ij}^{k} = \begin{cases} 0 & if \quad |x_{if_k} - x_{jf_k}| = min\left(|x_{if_k} - x_{lf_k}|\right) \\ 1 & if \quad |x_{if_k} - x_{jf_k}| > min\left(|x_{if_k} - x_{lf_k}|\right) \end{cases} \tag{2.5}$$

for each $l \in \{1, \ldots, N\} \setminus \{i\}$.

Dissimilarity for categorical features:

$$NNI_{ij}^{k} = \begin{cases} 0 & if \quad x_{if_k} = x_{jf_k} \\ 1 & if \quad x_{if_k} \neq x_{jf_k} \end{cases}, \tag{2.6}$$

where NNI_{ij}^{k} denotes the Nearest Neighbour Index for $\mathbf{X_i}$ and $\mathbf{X_j}$ individuals in the k-th dimension.

The Nearest Neighbour Index describing the dissimilarity of the two groups is calculated as the average of the dissimilarities calculated in each dimension.

$$NNI(X_T, X_{UT}) = \frac{\sum_{(\mathbf{X_i}, \mathbf{X_j}) \in M} \sum_{k=1}^{n} NNI_{ij}^{k}}{nN}, \tag{2.7}$$

where $(\mathbf{X_i}, \mathbf{X_j}) \in M$ yields that $\mathbf{X_i} \in X_T$ and $\mathbf{X_j} \in X_{UT}$ are matched case-control pairs.

As previously mentioned, the Nearest Neighbour Index only considers perfect correspondence, meaning, that the paired elements are closest or identical along the examined dimension, and it does not take into account the magnitude of difference. In contrast, the following measure aims to give a better understanding of the magnitude of difference of the paired elements, resulting in a more sophisticated and precise evaluation.

Global Dissimilarity Index

It is apparent that NNI checks for every dimension if the case-control pairs are closest to each other in that dimension, however, it does not consider the distance between them. The Global Dissimilarity Index (GDI) is a paired measure that is meant to account for this weakness.

GDI measures the dissimilarity for nominal features as the function of the number of different values, for ordinal features as the difference of ranks and for continuous features as the normalised distance. The statement about paired elements still holds.

Dissimilarity for continuous features:

$$GDI_{ij}^{k} = \frac{|x_{if_k} - x_{jf_k}|}{max_f - min_f}, \tag{2.8}$$

where GDI_{ij}^{k} denotes the Global Dissimilarity Index for $\mathbf{X_i}$ and $\mathbf{X_j}$ individuals in

the k-th dimension, min_f represents the minimum and max_f represents the maximum value measured in the f-th dimension taking into account all individuals from $X_T \cup X_C$.

Dissimilarity for nominal features:

$$GDI_{ij}^k = \begin{cases} 0 & if \quad x_{if_k} = x_{jf_k} \\ \frac{1}{F_k} & if \quad x_{if_k} \neq x_{jf_k} \end{cases}, \tag{2.9}$$

where F_k is the number of possible values along the k-th dimension.

Dissimilarity for ordinal features:

$$GDI_{ij}^k = \begin{cases} 0 & if \quad x_{if_k} = x_{jf_k} \\ \frac{|r_{x_{if_k}} - r_{x_{jf_k}}|}{F_k - 1} & if \quad x_{if_k} \neq x_{jf_k} \end{cases}, \tag{2.10}$$

where $r_{x_{if_k}}$ yields the ordered rank of the ordinal attribute x_{if_k} and $r_{x_{jf_k}}$ yields the ordered rank of the ordinal attribute x_{jf_k}.

The Global Dissmilarity Index describing the dissimilarity of the two groups is calculated as the average of the dissimilarities calculated in each dimension.

$$GDI(X_T, X_{UT}) = \frac{\sum_{(\mathbf{X_i}, \mathbf{X_j}) \in M} \sum_{k=1}^{n} GDI_{ij}^k}{nN}. \tag{2.11}$$

It can be seen that the Global Dissimilarity Index measures the magnitude of difference between the values of each dimension of the paired individuals not just identities if we found the identical value or the nearest neighbour along the examined dimension. Although both the Nearest Neighbour Index and the Global Dissimilarity Index are pairing-based dissimilarity measures, they approach the question of dissimilarity differently. The difference lies in the quality or magnitude of difference. NNI informs the analysts whether the pair chosen by their method is the closest possible individual from the given population, while GDI qualifies the actual aggregated difference between the paired individuals.

It is important to note, that in case of nearest neighbour-based pairing methods, it is possible that not the closest neighbour is chosen as a pair, but only the second or third nearest neighbour is selected. This can occur if the candidate individual is the closest neighbour for more than one individual from the case group, resulting in a conflict. So, to resolve this conflict, one of the conflicting individuals from the case group has to choose the next nearest neighbour, if appropriate. The NNI value

for such a case can drastically change, while the change in GDI is low, or in some extreme cases, non-existent. This shows that both measures behave differently but still contain valuable information regarding the dissimilarity of the examined case group and control group, however, it not recommended to use NNI as a standalone evaluation metric due to its crude nature.

2.2.1.2 Measure for distribution-based evaluation

The above-mentioned methods measure the dissimilarity by determining the pairwise dissimilarities for each case-control pair. However, not only the pairwise dissimilarities are relevant, but the similarities of the distributions of the characterising features have to be taken into account as well. For this reason, I suggested a distribution-based measure called Distribution Dissimilarity Index (DDI).

Distribution Dissimilarity Index

The Distribution Dissimilarity Index is based on the histogram disparities of the case group and the control group in each dimension. It overcomes the limitations of the widely used evaluation methods and is capable of handling all kinds of data including continuous, nominal and ordinal data. DDI relies on the absolute deviation of the frequency of each property value relative to the size of the control group and the number of characterising features. If the individuals are characterised by continuous values, the values have to be discretised before the calculation of the frequency values. Because the method of discretisation (equal width binning, equal frequency binning, other binning methods) significantly affects the dissimilarity results, the most suitable discretisation method for the field of the study is recommended for use (e.g., the age attribute can be discretised based on the population pyramid in case of healthcare).

After data preparation, the Distribution Dissimilarity Index is calculated the following way:

$$DDI(X_T, X_{UT}) = \frac{\sum_k \sum_{v \in V_k} |freq_{kv}^{X_T} - freq_{kv}^{X_{UT}}|}{nN}, \tag{2.12}$$

where $freq_{kv}^{X_T}$ yields the absolute frequency of the v-th category in the k-th dimension in the case group, $freq_{kv}^{X_{UT}}$ analogously for the control group, V_k is the set of possible categories along the k-th dimension, and $k = 1, ..., n$.

We can see, that this general, frequency-based measure does not assume any distribution along any of the dimensions, as opposed to the Student's t-test, making

it adequate to use without any restraints in case of multidimensional data.

It is important to note, that the dissimilarity value of 0 has different meaning in the two cases. While in case of the paired evaluations the dissimilarity value 0 means that the cohorts X_T and X_{UT} contain the same individuals from the viewpoint of pairing, in the non-paired case it yields only the identity of the distributions of the dimensions. The 1 value in both cases indicates that the two groups differ as much as possible from each other.

The main advantages of the three proposed dissimilarity measures are that they express the dissimilarity of the case group and the control group as a single dissimilarity value in the range of $[0, 1]$ as opposed to the multiple values of the methods presented in Section 2.1.2. This dissimilarity value can be calculated for multidimensional data, regardless of the types (e.g., ordinal, nominal continuous) of the attributes.

2.2.2 Evaluation of the proposed measures

Before application, the proposed measure had to be evaluated. The evaluation of the proposed measures was performed in three different ways. First, the responsiveness of the proposed measures was evaluated, the second examination compared different candidate control groups by the use of the proposed similarity measures, and the third one aimed to evaluate the behaviour of the measures in a quasi-real scenario. All evaluations were performed by Monte Carlo simulations and are described in detail in the following subsections.

2.2.2.1 Data generation

Evaluating the suggested measures requires reliable data sources. Real healthcare datasets do not always hold this criteria or are not publicly available. Thus, a dataset-generator was implemented to provide a controlled environment for making deductions. The dataset-generator was implemented in Python and it is capable of generating the following data types (in any possible combination): binary Bernoulli random variable, binomial variable, continuous variable with normal distribution (by mean and variance, or in range), continuous variable with uniform distribution, and discrete variable with quasi-uniform distribution. Furthermore, the developed Monte Carlo simulator is able to execute predefined (research-dependent) operations on the generated dataset: adding noise to the data, calculating measures and evaluating results.

Using this simulator it was possible to model any healthcare-related or other arbitrary datasets, eliminating the need to use unreliable data sources. As a result, evaluations happened in a controlled, well-defined environment.

2.2.2.2 Responsiveness of the measures to noise

Healthcare related data is usually noisy. To demonstrate the responsiveness of the measures to noise, a complex dataset containing all variable types was generated. The generated dataset contained a 1000 elements (representing the individuals) and element was characterised by 8 variables (2 binary, 2 ordinal, 1 nominal and 3 continuous): 1 Bernoulli random variable with a probability value of 0.5 ($\sim$B$(1, 0.5)$), 1 Bernoulli random variable with a probability value of 0.3 ($\sim$B$(1, 0.3)$), 1 binomial variable with 3 trials and a probability value of 0.5 ($\sim$B$(3, 0.5)$), 1 uniform discrete variable in the range of $[0, 5)$ ($\sim$U$_{disc}(0, 5)$), a uniform discrete variable in the range of $[0, 4)$ ($\sim$U$_{disc}(0, 4)$), 1 uniform variable in the range of $[0, 2)$ ($\sim$U$(0, 2)$), 1 variable with normal distribution with a mean of 2 and standard deviation of 0.5 ($\sim\mathcal{N}(2, 0.5)$) and 1 variable with normal distribution with a mean of 1 and standard deviation of 2 ($\sim\mathcal{N}(1, 2)$).

The original dataset was distorted with different degrees of noise: $1\,\%$, $5\,\%$, $10\,\%$, $25\,\%$, $50\,\%$, $75\,\%$, $90\,\%$, and finally $100\,\%$ of the dataset was distorted with noise along each dimension. The added noise was attribute-dependent: for binary variables the values were negated, for nominal and ordinal variables the values were shifted and aligned for the range, and for continuous variables the value was changed by at most $\pm 10\,\%$. My goals was to achieve theoretical completeness, but it is important to note that noise levels of $50\,\%$ and above are not realistic in medical datasets. From a practical point of view, the results connected to lower noise levels are more relevant.

In total, 9 case-control group pairs were created to test the proposed measures. In each case, the original, noiseless dataset and a noisy dataset formed a pair. The quality of the case-control group pairs was evaluated by averaging the NNI, the GDI, and the DDI values of a 100 runs. The results of the evaluation can be seen in Table 2.1.

The presented values in Table 2.1 are dissimilarities in the range of $[0, 1]$. The smaller the value, the more similar the given case and control groups are. It can be seen that in case of identical case and control group pairs (noise is $0\,\%$), the value of all proposed measures was 0, and by increasing the amount of noise, the

Table 2.1. Average of NNI, GDI and DDI values for the case-control group pairs, where the control groups were distorted with different levels of noise.

		Dissimilarity measure		
	Noise	NNI	GDI	DDI
Noisy control	0 %	0.000	0.000	0.000
	1 %	0.009	0.004	0.003
	5 %	0.043	0.019	0.008
	10 %	0.085	0.038	0.015
	25 %	0.214	0.096	0.036
	50 %	0.426	0.192	0.054
	75 %	0.645	0.291	0.080
	90 %	0.772	0.347	0.091
	100 %	0.851	0.385	0.097

dissimilarity values also increase for all three measures. However, the magnitude of the change differs. The change is the largest in the case of NNI. As previously mentioned, NNI is the strictest measure, so it is especially sensitive to the noise and to dissimilar data. The 0.851 dissimilarity value reinforces the previous statement about the behaviour of NNI. The statement about strict nature also holds for GDI, while DDI, the non-paired measure is noticeably less sensitive to noise, reaching only 0.097 dissimilarity value when the whole dataset is distorted.

It is important to mention that total dissimilarity (when the dissimilarity value is 1) is only achievable in extreme cases. These extreme cases are where the compared values are at the opposite ends of the range of the examined variable.

Figure 2.1 shows the values of all measures in function of the distortion. We can seen that it is linear in case of all measures. The gradient of the lines is different as the responsiveness of the measures is different as well. The values of the proposed measures should not be directly compared as they evaluate the dissimilarity from different aspects, they should be used for evaluation in conjunction.

2.2.2.3 Comparison of possible control groups

The aim of the second evaluation was to compare the proposed measures and examine their behaviour for different candidate control groups. For this purpose, a dataset containing 2000 elements was generated. All individuals were charac-

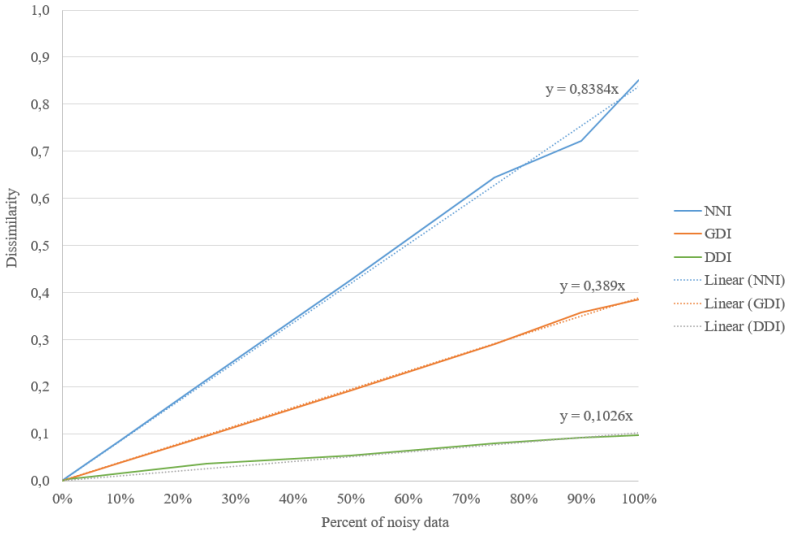

Fig 2.1. The average NNI, GDI and DDI values in function of distortion.

terised by 6 variables: 2 Bernoulli random variables with a probability value of 0.5 ($\sim$B$(1, 0.5)$), 1 ordinal variable in the range of $[0, 5)$ ($\sim$B$(6, 0.5)$), 1 ordinal variable in the range of $[0, 4)$ ($\sim$B$(4, 5)$), 1 uniform variable in the range of $[0, 5)$ ($\sim$U$(0, 5)$) and 1 variable with normal distribution with a mean of 2 and standard deviation of 0.5 ($\sim\mathcal{N}(2, 0.5)$).

A case group with size 150 was randomly selected from the previously generated population. In addition, 8 control groups with a size of 150 individuals were selected also randomly from the remaining population. The original dataset was not distorted with noise, so groups selected by random sampling define 8 different control groups. The created 8 case-control group pairs (the first one as a sample and the 8 others as possible controls) comprised my test scenario. The dissimilarities of all pairs were evaluated by NNI, GDI, and DDI. The results can be seen in Table 2.2.

The presented values in Table 2.2 are dissimilarities in the range of $[0, 1]$. The smaller the value, the more similar the given case and control groups are. Cells highlighted with light grey colour contain the minimal value for each measure. In accordance with the used measure, these control groups are the most similar (best match) to the case group.

It can be seen that the approach of the similarity evaluation greatly affects the selection of the best control group. Not only the magnitudes of the measures differ but also the order of the control groups in terms of similarity. For NNI the 4th, for GDI the 7th, and for DDI the 3rd and 5th control groups are the most similar control groups. The least similar control groups are also different: the 5th for NNI, the 6th

Table 2.2. NNI, GDI, DDI and overall dissimilarity for the 8 case-control group pairs.

		Dissimilarity measure			
	No.	NNI	GDI	DDI	overall
Control	#1	0.761	0.422	0.061	0.450
	#2	0.757	0.409	0.072	0.463
	#3	0.762	0.414	0.055	0.285
	#4	0.749	0.407	0.068	0.314
	#5	0.789	0.426	0.055	0.627
	#6	0.783	0.430	0.056	0.633
	#7	0.758	0.396	0.068	0.281
	#8	0.772	0.424	0.076	0.780

for GDI and the 8th for DDI. It is interesting to notice that one of the most similar control group for DDI (control group #5) is the least similar for NNI. This can be accounted to the fact that all proposed similarity measures determine similarity from different aspects, as it is described in Section 2.2.1.

These observations raise the need for a complex, all-encompassing evaluation. Considering the 2 pairing-based measures (NNI and GDI) and calculating the normalised average value of them, the 7th control group seems to be the appropriate choice, however if DDI is also taken into account the 3rd control group is not a bad choice either. This fact is also confirmed by the overall dissimilarity measure, which was calculated as the average of the individually normalised dissimilarity measures (see the last column in Table 2.2).

My results highlighted that the evaluation of case and control groups always requires more than one measure, as different measures give different insights regarding the quality of the selected control group. Unfortunately, in most of the published articles in the literature at most one measure is used, which is, in my opinion, insufficient.

2.2.2.4 Application of the measures in case of Propensity Score Matching

To test the applicability of the proposed dissimilarity measures, a real-life application scenario was modelled. Subjects of the dataset were characterised by 10 baseline binary covariates and a treatment status indicator (Z_i) was determined for each

subject. Subjects with $Z_i = 1$ were considered treated and subjects with $Z_i = 0$ were considered untreated patients. Based on this model, a population containing 1000 subjects was generated. The generated data was separated on the status of treatment. Treated subjects ($Z_i = 1$) composed the case group and untreated subjects ($Z_i = 0$) composed the population of the possible controls. The 1000 subjects were separated in a $30 - 70\,\%$ ratio: $30\,\%$ of the subjects was considered as members of the treated case group and the other $70\,\%$ was considered the untreated population from which a 100 different control groups were selected with propensity score matching. During the propensity score matching a logistic regression classification model was determined to estimate the effect of the baseline covariates to the exposure of the treatment. Based on the logistic regression model the propensity scores for each individual were calculated, and these estimated values served as the basis of pairing elements between the treated and untreated groups.

The quality of the propensity score-based control group selection method can be controlled by the caliper size parameter of the PSM algorithm. The caliper size suggested in the literature for PSM can be calculated as described in [21].

$$0.2\sqrt{(\sigma_T^2 + \sigma_{UT}^2+)/2}, \tag{2.13}$$

where σ_T is variance of the covariates of X_T and σ_{UT} is variance of the covariates X_{UT}.

For testing the quality of the control groups selected with different caliper size parameters, different multipliers (0.5, 0.75, 1.00, 1.25, 1.50, 2.00, 2.50, 5.00 and 10.0) were determined for the caliper size presented in Eq. 2.13.

Figure 2.2 shows the relation between the size of the selected control group and the caliper size multiplier. The suggested caliper size suggested is marked with the 1.00 multiplier and a darker colour. It can be seen that by decreasing the caliper size, the size of the selected control group also decreases. By increasing the caliper size, further and further individuals can be selected into the control group, which in turn, decreases the quality of the control group. This decrease in quality can be measured by the measures proposed in Section 2.2.1.

In Figure 2.3 and Figure 2.4 the relation between the proposed NNI and GDI measures for paired evaluation and the caliper size multiplier can be seen. The figures show that by increasing the caliper size the similarity of the control group decreases, which can be seen in the increase of the proposed dissimilarity measure values. The larger the caliper size, the further subjects can be paired from the pop-

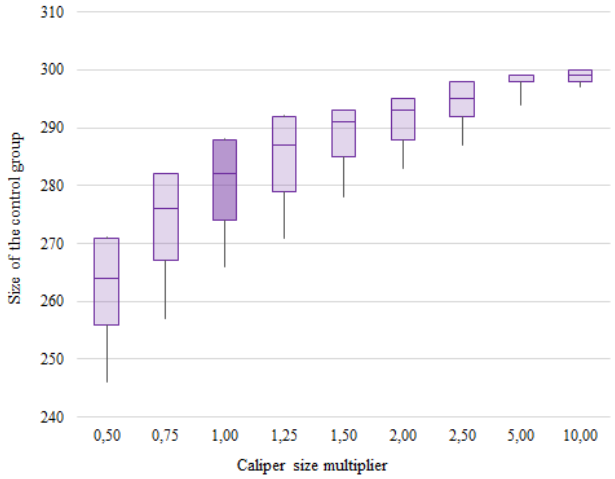

Fig 2.2. The size of the control group in function of the caliper size multiplier.

ulation to the subjects of the case group. That is, the quality of the control group is constantly decreasing.

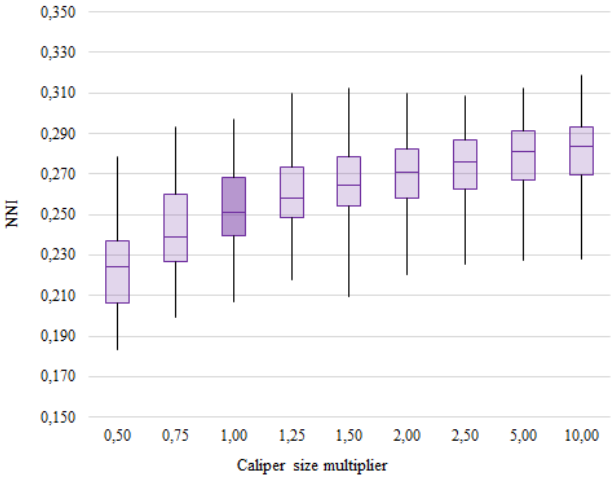

Fig 2.3. NNI in function of the caliper size multiplier.

Finally, Figure 2.5 shows the relation between DDI and the caliper size multiplier. There is no noticeable effect of changing the caliper size on the DDI measure. The cause of this is that DDI is a distribution-based measure and all control groups were selected from the same population, meaning that their distributions are also similar to the distribution of the case group. It is no surprise, that this measure does not reflect the increase of the pairwise dissimilarities, however, it reinforces the fact that the similarity of case and control groups needs to be evaluated from different aspects as well.

After reviewing the proposed dissimilarity measures and their characteristics, I deal with the problems of control group selection in the next section.

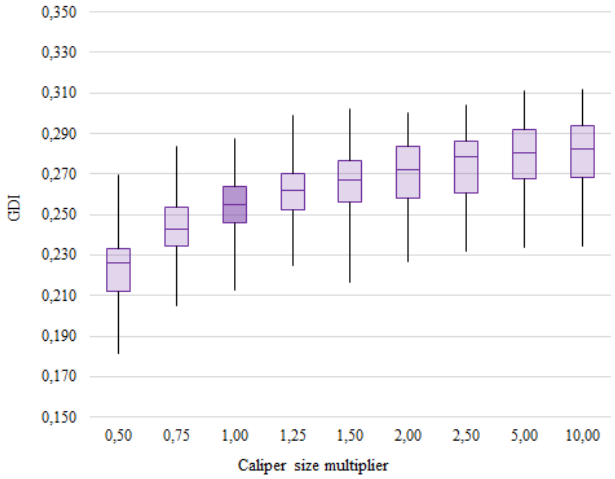

Fig 2.4. GDI in function of the caliper size multiplier.

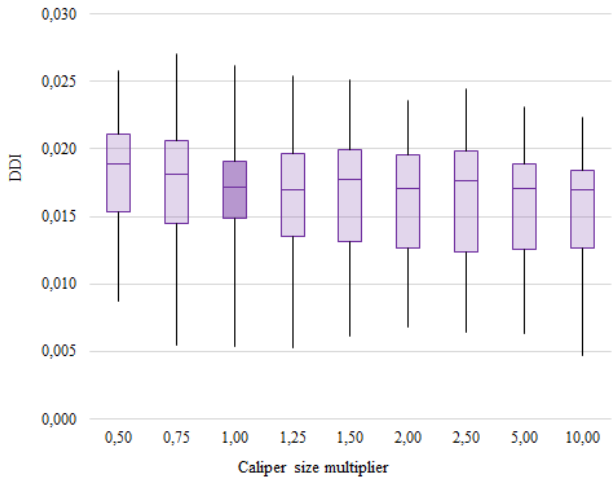

Fig 2.5. DDI in function of the caliper size multiplier.

2.3 Novel control group selection methods

As it was described in Section 2.1.1.1, PSM methods are constantly in the midst of criticism, due to the imbalance of the covariates observed in some studies.

The disadvantage of the most widely used PSM and all other propensity score-based methods is that they perform the matching of individuals in the 1-dimensional space of the propensity scores. Additionally, uncertainty is also increased by the fact that propensity scores are estimated and not known a priori. Although some methods have been proposed earlier to pair individuals in the original vector space of the features [93, 94, 95], to the best of my knowledge, none of them utilises the result of fitting a logistic regression model during the matching performed in the original vector space. Taking advantage of this opportunity, I developed a novel

control group selection method which not only matches individuals in the original, n-dimensional space, but also weights each dimension by its significance which is determined by the the odds ratio values of the logistic regression fit. The proposed method is named *Weighted Nearest Neighbours Control Group Selection with Error Minimization*.

2.3.1 Weighted Nearest Neighbours Control Group Selection with Error Minimization

The suggested Weighted Nearest Neighbours Control Group Selection with Error Minimization (WNNEM) method [42] considers each subject as an n-dimensional data point in an n-dimensional space, where each covariate ($f_k, k = 1, \ldots, n$) represents a unique dimension. This way, the problem of control group selection can be interpreted as a distance minimisation problem. To select a proper control group, such individuals have to be identified from the candidates that lie close to the individuals of the case group. The concept of lying close can be defined in numerous ways. In the case of the proposed method, multivariate matching is performed in which the odds ratio (OR) values of the fitted logistic regression model are utilised as weighting factors of the covariates to compute the distances between the individuals.

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

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

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

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

$$d_{ij}^{(f)} = \begin{cases} 0 & \text{if} x_{if} = x_{jf} \\ 1 & \text{otherwise} \end{cases}, \tag{2.14}$$

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

- In case of nominal variables simple matching distance (Eq. 2.14) is calculated. Incidentally, these variables can be coded as a set of binary variables and the distance can be calculated as the normalised distance of binary features, where the normalisation constant is the number of possible values of the nominal variable.

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

$$x_{if}^{*} = \frac{x_{if} - min_f}{max_f - min_f}, \tag{2.15}$$

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

$$d_{ij}^{(f)} = |x_{if}^{*} - x_{jf}^{*}|, \tag{2.16}$$

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

After ensuring that the meaning of the dissimilarity values is identical in each dimension, the next step is to weight them according to their relevance to group (treatment) assignment.

As previously mentioned, the odds ratio values of the logistic regression model fitted on the status of treatment assignment are utilised for this purpose. Generally, the odds ratio is the probability of a characteristic being present divided by the prob-

ability of the same characteristic being absent. The odds ratio for each independent variable can be obtained by applying the exponential function to the corresponding regression coefficient (b_i) obtained from the logistic regression model described by Eq. 2.2.

Odds ratios as the weights of the covariates are calculated as

$$w_i = OR_i = e^{b_i}, \tag{2.17}$$

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

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

$$dist(\mathbf{X}_i, \mathbf{X}_j) = \sum_{f=1}^{n} w_f d_{ij}^{(f)}, \tag{2.18}$$

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

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

My practical experiments showed that for 1:1 matching, an adequate solution can be found even without the use of a complex optimisation algorithm. The only problem that needs to be handled during optimisation is how to manage the pairing process of those candidates which lie closest to more than one individual from the case group. These candidates are called *candidates in conflict* and are formally defined as follows: $\mathbf{X}_j \in X_C$ is a *candidate in conflict* if $dist(\mathbf{X}_i, \mathbf{X}_j)$ is minimal for more than one $\mathbf{X}_i \in X_T$. For handling these conflicts, the order of the neighbours has to be determined.

Let $NN_1(\mathbf{X}_i)$ denote the closest and $NN_2(\mathbf{X}_i)$ the second closest neighbour to individual $\mathbf{X}_i \in X_T$.

$$NN_1(\mathbf{X}_i) = \underset{\mathbf{X}_j \in X_C}{\operatorname{argmin}}(dist(\mathbf{X}_i, \mathbf{X}_j)). \tag{2.19}$$

$$NN_2(\mathbf{X}_i) = \underset{\mathbf{X}_j \in X_C - \{NN_1(\mathbf{X}_i)\}}{\operatorname{argmin}}(dist(\mathbf{X}_i, \mathbf{X}_j)). \tag{2.20}$$

The design of the conflict-handling method to solve the competition of two indi-

viduals was inspired by the Vogel-Korda method (Vogel's Approximation Method, VAM): instead of a greedy selection, the second neighbours of the treated individuals are also taken into account: the candidate in conflict is matched to the individual for which the error function is greater. The error function is calculated as the distance of the first and second neighbours of the individuals.

$$E(\mathbf{X}_i) = |dist(\mathbf{X}_i, NN_1(\mathbf{X}_i)) - dist(\mathbf{X}_i, NN_2(\mathbf{X}_i))|. \tag{2.21}$$

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

The steps of the proposed Weighted Nearest Neighbours Control Group Selection with Error Minimization method (WNNEM) are summarised by Algorithm 2.1. We assume that $argmax$ returns only 1 index.

2.3.2 Evaluation of the proposed WNNEM method

To present the effectiveness of the proposed method, a Monte Carlo simulation-based evaluation was performed. During the evaluation, the quality of the control group resulting from the proposed WNNEM method was compared to the quality of the control group selected by the most widely applied form of the PSM method, namely, to the result of the greedy 1:1 propensity score matching performed without replacement of individuals and utilising a proper caliper size for the selection procedure. The proper caliper size in each simulation was determined dynamically and was set at the minimal value for which 1:1 matching could be performed.

2.3.2.1 Datasets

For the comparisons, three scenarios with varying feature characteristics were designed. For each scenario, 20 datasets were generated randomly with the same distribution parameters predefined for the covariates. As a result, each scenario contained 20 individual datasets with the same number of individuals.

I used a benchmark dataset widely applied in theoretical PSM studies in Scenario I [96]. Scenario II and Scenario III simulated such studies in which the age of the patients and another 5 binary parameters were considered as covariates. With

Algorithm 2.1: Weighted Nearest Neighbours Control Group Selection with Error Minimization (WNNEM)

Input: X_T case group, X_C set of candidate individuals

Output: X_{UT} control group

1 Perform a logistic regression to estimate w_i weights for all covariates.

2 Normalise X_T and X_C collectively using feature scaling and calculate the $\mathbf{D} = dist(\mathbf{X}_i, \mathbf{X}_j)$ distance matrix for all pairs of individuals of $\mathbf{X}_i \in X_T$ and $\mathbf{X}_j \in X_C$ by Eq. 2.18.

3 Set

$$X_{unpaired} = X_T$$
$$X_{UT} = \emptyset$$

4 Determine $NN_1(\mathbf{X}_i)$ and $NN_2(\mathbf{X}_i)$ based on the distance matrix $\mathbf{D}$ for all $\mathbf{X}_i \in X_{unpaired}$.

5 Calculate $E(\mathbf{X}_i)$ for all $\mathbf{X}_i \in X_{unpaired}$.

6 For $i = 1, \ldots, \|X_{unpaired}\|$

Set $k = \text{argmax}_{\mathbf{X}_i \in X_{unpaired}}(E(\mathbf{X}_i))$

If $NN_1(\mathbf{X}_k) \notin X_{UT}$:

$X_{UT} = X_{UT} \cup \{NN_1(\mathbf{X}_k)\}$

$X_{unpaired} = X_{unpaired} - \{\mathbf{X}_k\}$

Set $l = arg(NN_1(\mathbf{X}_k))$

Set $\mathbf{D}(i, l) = \infty$ for all $i = 1, \ldots, \|X_T\|$

7 Repeat *Steps 4 to 6*, till $X_{unpaired} \neq \emptyset$.

the analysis of theses scenarios, I aimed to simulate such recurring biomedical studies which are based on a few covariates that are mainly binary. In Scenario II, the size of the treated group varied between $16\,\%$ and $20\,\%$ of the dataset, and in the case of Scenario III, it was between $24-27\,\%$. In other words, Scenario III simulated a more difficult case, where the ratio of the candidate individuals to the treated subjects was lower ($X_C/X_T \approx [1.5, 2.2]$) than in Scenario II ($X_C/X_T \approx [2.4, 3.2]$), therefore, it was harder to find a proper pair for each treated individual.

As PSM is non-deterministic and dependent on the order of the individuals it was performed 5 times on each of the 20 generated datasets and the matching order in each experiment was randomised during the simulations. When the WNNEM method was applied, because of the deterministic nature of the algorithm, the control group selection was performed only once for each of the 20 generated datasets.

2.3.2.2 Scenario I

Scenario I is a widely used benchmark dataset and all parameters are taken from [96]. According to this dataset, all individuals are characterised by 10 binary variables, each from a Bernoulli distribution ($x_j \sim \mathrm{B}(0.5), \quad j = 1, \ldots, 10$). To calculate the probability of treatment assignment, the following logistic regression model was used.

$$\begin{aligned} \mathrm{logit}(p_{i,treat}) = b_{0,treat} + \\ b_L x_{i1} + b_L x_{i2} + b_L x_{i3} + b_M x_{i4} + b_M x_{i5} + \\ b_M x_{i6} + b_H x_{i7} + b_H x_{i8} + b_{VH} x_{i9} + b_{VH} x_{i10} \end{aligned} \quad . \tag{2.22}$$

A treatment status indicator (Z_i) was generated for each subject from a Bernoulli distribution with a subject-specified probability equal to $p_{i,treat}$ ($Z_i \sim \mathrm{B}(p_{i,treat})$). The treated group consisted of subjects where $Z_i = 1$, while subjects where $Z_i = 0$ were assigned to the untreated group (from which the control group was selected). The b weights in Eq. 2.22 denote a low (L), medium (M), high (H) or very high (VH) effect on treatment assignment. The weights were assigned in such a way that approximately $25\,\%$ of the subjects were treated. The number of individuals in each dataset was 1000. The applied weight coefficients were as follows:

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

- high: $b_H = \log(1.5)$
- very high: $b_{VH} = \log(2.1)$

20 independent datasets were generated with the previously described parameters. PSM was executed 5 times and WNNEM only once for each dataset. These simulations resulted in 100 control group selections for the PSM method and 20 control group selections for the WNNEM method.

Table 2.3 summarises the average, minimum, and maximum p-values for the Hansen and Bowers imbalance test, as well as the minimum, average and maximum distance measures for the DDI, NNI and GDI dissimilarity measures. I decided to use the Hansen and Bowers test, as it is applied for more complex evaluations in biomedical studies. This measure allows the evaluation of the imbalance of all covariates simultaneously: covariates are considered poorly balanced if the test value is significant ($p < 0.05$). The higher the p-value, the more similar the case and control groups are. In case of the dissimilarity measures (DDI, NNI, and GDI), the dissimilarity values falls within the range of $[0, 1]$, but the value of zero expresses that the case and control groups are identical. Consequently, the greater the dissimilarity value, the higher the difference of the case and control groups is.

Table 2.3. Quality measures for Scenario I.

	PSM			WNNEM		
	min	avg	max	min	avg	max
HB(p)	0.722	0.960	1.000	0.967	0.996	1.000
DDI(d)	0.010	0.017	0.026	0.006	0.011	0.018
NNI(d)	0.215	0.280	0.344	0.053	0.060	0.067
GDI(d)	0.220	0.314	0.397	0.051	0.061	0.074

Results presented in Table 2.3 show that the WNNEM method performed better in terms of control group selection than the greedy 1:1 PSM. All dissimilarity values, both for the paired evaluations (NNI and GDI) and the distribution-based evaluation (DDI), are lower in the case of control groups selected by the WNNEM method than by greedy PSM. Furthermore, the p-value of the Hansen and Bowers test also shows a higher degree of balance when the WNNEM method was applied.

For a detailed insight, Table 2.4 presents the p-values of the Hansen and Bowers imbalance test for each dataset. $|X_T|$ yields the number of the treated individuals. In the case of the PSM method, the minimum, average and maximum values for the 5 experiments are shown. As for each dataset, the WNNEM method was run

only once, therefore only one p-value is given. The differences of the p-values (diff(p)) were calculated as the average p-value of the PSM method subtracted from the p-value of the WNNEM method. The p-values for all PSM simulations and the WNNEM method are also presented in Figure 2.6.

Table 2.4. Results of the Hansen and Bowers test in Scenario I.

		PSM			WNNEM	
dataset	$\|X_T\|$	$min(p)$	$avg(p)$	$max(p)$	p	diff(p)
1	337	0.961	0.972	0.986	0.999	0.027
2	321	0.916	0.950	0.972	0.967	0.017
3	316	0.993	0.996	0.998	1.000	0.004
4	341	0.995	0.998	1.000	0.978	-0.020
5	354	0.990	0.995	0.998	0.998	0.003
6	335	0.994	0.996	0.998	1.000	0.004
7	317	0.995	0.997	0.998	1.000	0.003
8	320	0.941	0.960	0.973	1.000	0.040
9	325	0.998	0.999	1.000	0.998	-0.001
10	317	0.975	0.986	0.993	0.997	0.011
11	315	0.834	0.878	0.898	1.000	0.122
12	319	0.969	0.988	0.997	0.999	0.011
13	287	0.998	0.999	1.000	1.000	0.001
14	329	0.887	0.901	0.916	0.999	0.098
15	338	0.772	0.811	0.876	1.000	0.189
16	344	0.887	0.911	0.935	0.986	0.075
17	309	0.995	0.998	0.999	0.999	0.001
18	335	0.907	0.936	0.959	1.000	0.064
19	325	0.919	0.940	0.971	0.997	0.057
20	335	0.976	0.981	0.991	1.000	0.019
min						-0.020
avg						0.036
max						0.189
sum						0.725

As can be seen in Table 2.4 and Figure 2.6, considering the simulated 20 datasets, on average, the greedy PSM method produced a better control group than the proposed WNNEM method only twice, namely for dataset 4 and dataset 9. The difference between the p-values in both cases is marginal. However, the WNNEM method in many cases (see datasets 8, 11, 14-16, 18 and 19) resulted in a much more similar control group to the group of treated individuals than the control groups selected by the PSM method. This fact is also shown numerically by the aggregated statistics at the bottom of Table 2.4 (see the minimum and maximum of the difference). This

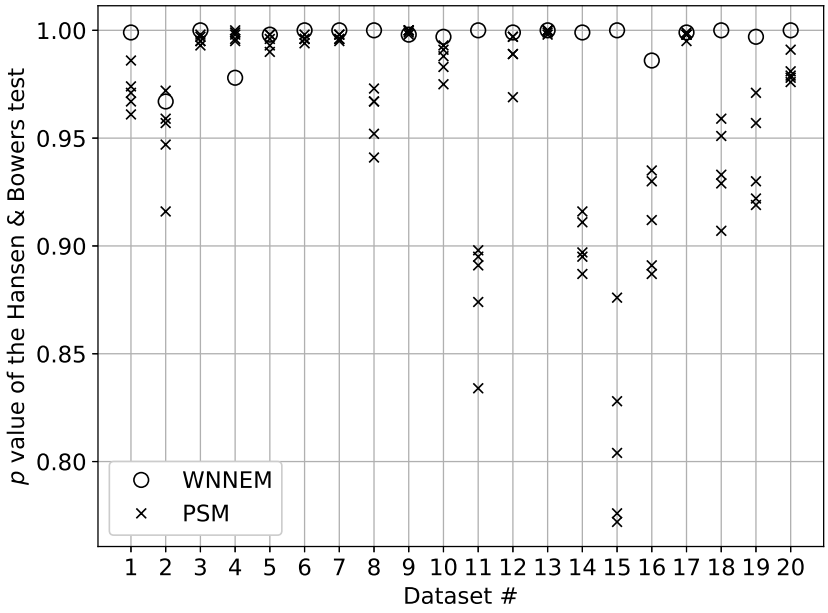

Fig 2.6. Results of the Hansen and Bowers test for each dataset in Scenario I.

part of the table also contains the totalled differences of the p-values for all datasets, which value also confirms the advantage of applying the WNNEM method. Table 2.4 also highlights that the WNNEM method identified a perfectly balanced control group (p-value is equal to 1) for 9 datasets, while the PSM method could for only 3 datasets. Furthermore, it must not be forgotten, that the WNNEM method was run only once, while the PSM method was run 5 times on each dataset.

To evaluate the results of the selected control groups, the similarity of the covariates were also evaluated separately. In this regard, for each control group selection, the similarity of the values of the covariates for the case and control groups was tested by the Chi-square test. A higher p-value means a more balanced control group in terms of a given covariate. The detailed results are presented as box plots in Figure 2.7. As can be seen, the median of the p-values for each covariate is higher in the case of WNNEM, and the interquartile range is also smaller for all covariates.

To sum up, we have shown in this subsection that the proposed WNNEM method may provide better results than the greedy, 1:1 PSM method on the benchmark dataset.

2.3.2.3 Scenario II

Scenario II models such studies in which fewer descriptive variables are available. In this scenario, each individual is characterised by 1 ordinal and 5 binary variables. The ordinal variable represents 5 age groups, while the binary variables may

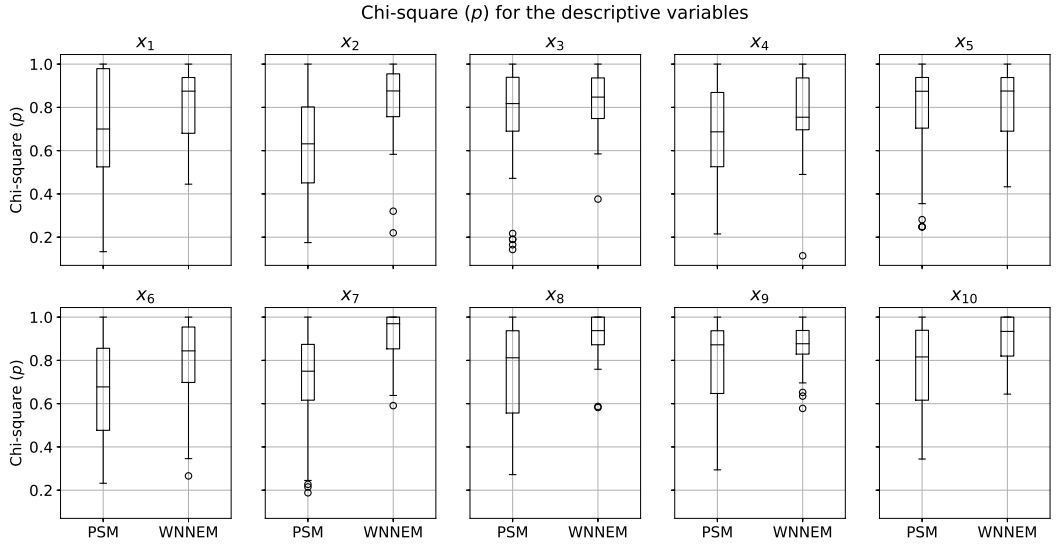

Fig 2.7. Distribution of all covariates in Scenario I.

represent, for example, the gender of the subject or various diagnoses.

The assignment of treatment status was analogous to the previously described assignment. In this scenario, the assignment of weights to each descriptive variable was as follows: the ordinal variable (x_1) had very high effect, as it is usual for age, the binary variables had low (x_2), medium ($x_3 - x_5$) and high (x_6) effect on the status of treatment. The ratio of the candidate subjects to the treated individuals in the 20 datasets was between 2.4 and 3.2.

The overall statistics of the control group selections are presented in Table 2.5. It can be seen that the proposed WNNEM method also performed better in this scenario. All dissimilarity measures exhibited lower degrees of dissimilarity and the p-value of the Hansen and Bowers test also exhibited a higher degree of balance overall. Furthermore, by comparing these results to the results of Scenario I, it can be seen that the selected control groups are more similar to the subjects of the case group.

Table 2.5. Quality measures for Scenario II.

	PSM			WNNEM		
	min	avg	max	min	avg	max
HB(p)	0.977	0.998	1.000	0.995	1.000	1.000
DDI(d)	0.002	0.008	0.019	0.000	0.003	0.009
NNI(d)	0.002	0.012	0.028	0.001	0.006	0.011
GDI(d)	0.002	0.013	0.310	0.001	0.005	0.011

Table 2.6 and Figure 2.8 present the detailed results for each of the 20 simulated

datasets. As it can be seen, both methods were able to select a perfectly balanced control group in cases of datasets 3-5, 8-10, 12-16, 18 and 19. This is due to the lower variability of features of the individuals and the relatively large number of available candidates. In dataset 1, 2, 7 and 20, the PSM method could not select a perfectly balanced control group, but the WNNEM method could. The opposite was true for dataset 11, however, in this case, the difference between the p-values was marginal.

Table 2.6. Results of the Hansen and Bowers test in Scenario II.

		PSM			WNNEM	
dataset	$\|X_T\|$	$min(p)$	$avg(p)$	$max(p)$	p	diff.(p)
1	195	0.999	0.999	0.999	1.000	0.001
2	183	0.996	0.996	0.996	1.000	0.004
3	190	1.000	1.000	1.000	1.000	0.000
4	197	1.000	1.000	1.000	1.000	0.000
5	167	1.000	1.000	1.000	1.000	0.000
6	201	0.977	0.977	0.977	0.999	0.022
7	196	0.999	0.999	0.999	1.000	0.001
8	198	1.000	1.000	1.000	1.000	0.000
9	185	1.000	1.000	1.000	1.000	0.000
10	196	1.000	1.000	1.000	1.000	0.000
11	183	0.999	1.000	1.000	0.998	-0.002
12	182	1.000	1.000	1.000	1.000	0.000
13	176	1.000	1.000	1.000	1.000	0.000
14	202	1.000	1.000	1.000	1.000	0.000
15	191	1.000	1.000	1.000	1.000	0.000
16	168	1.000	1.000	1.000	1.000	0.000
17	179	0.997	0.997	0.997	0.995	-0.002
18	185	1.000	1.000	1.000	1.000	0.000
19	204	1.000	1.000	1.000	1.000	0.000
20	201	0.983	0.983	0.985	1.000	0.017
min						-0.002
avg						0.002
max						0.022
sum						0.041

For further evaluation, the similarity of the covariates was also calculated. The box plots (Figure 2.9) show that the WNNEM method was able to select more similar control groups than the greedy 1:1 PSM method for every covariate. It is important to emphasize that in the case of completely missing boxes (for covariates x_1, x_3, x_4 and x_6 for the WNNEM method), the first, second and third quartiles of

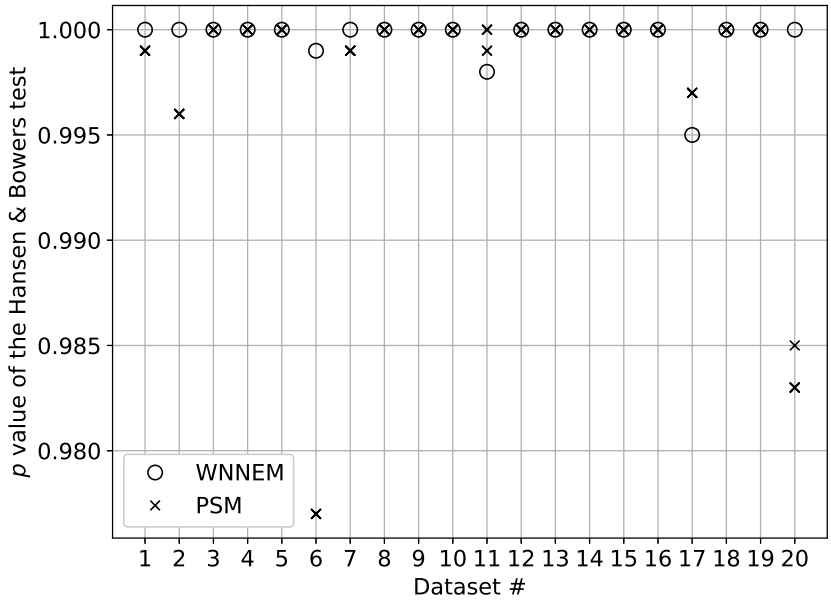

Fig 2.8. Results of the Hansen and Bowers test for each dataset in Scenario II.

the p-values were all equal to 1. In the case of partially missing boxes, the median was equal to 1, therefore the third quartile and maximum value were equal. Figure 2.9 shows that the WNNEM method achieved perfect matching on the most important covariates (x_1 and x_6) while the applied PSM method could not. Furthermore, in the case of PSM, the largest imbalance was observed for a covariate of medium effect (x_4), while by applying the proposed WNNEM method, it was observed for a covariate of low effect (x_2).

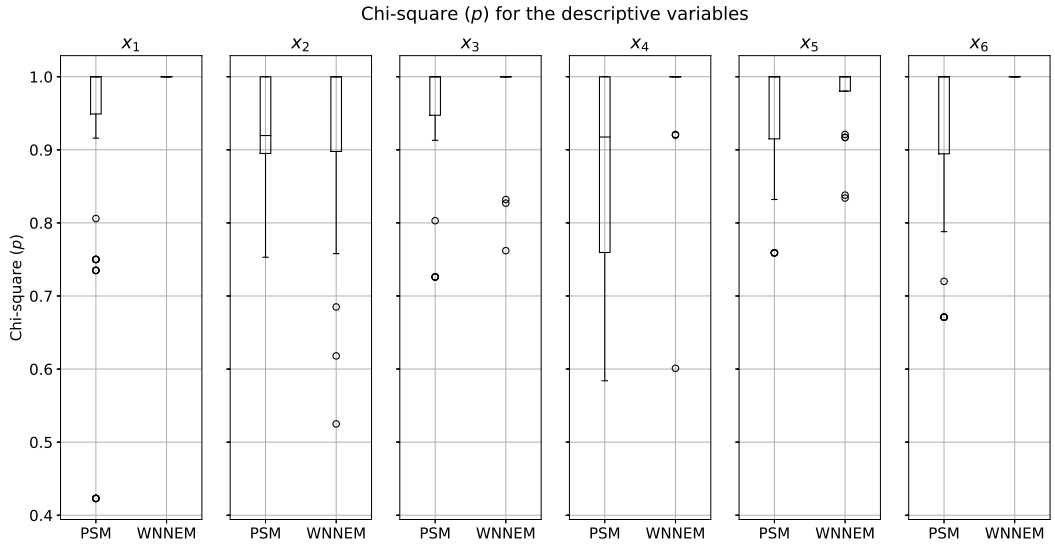

Fig 2.9. Distribution of all covariates in Scenario II.

2.3.2.4 Scenario III

Scenario III is similar to Scenario II with regard to the attributes of individuals and the total number of subjects in each dataset. However, it simulates a more difficult control group selection problem. The number of treated individuals in the case of the third scenario is higher than in the second one. In the third scenario, the ratio of the candidate individuals to the treated ones was only between 1.5 and 2.2, thus, on average, only 2 candidate individuals were available per treated person.

In Table 2.7, the overall dissimilarity measures and the results of the Hansen and Bowers tests are presented. By comparing Tables 2.5 and 2.7, it can be seen that in the case of the third scenario, it was harder to select a fully balanced control group using both methods. However, Table 2.7 shows that the WNNEM method was able to select more balanced control groups than the greedy 1:1 PSM method.

Table 2.7. Quality measures for Scenario III.

	PSM			WNNEM		
	min	avg	max	min	avg	max
HB(p)	0.743	0.954	1.000	0.896	0.982	1.000
DDI(d)	0.005	0.015	0.030	0.003	0.010	0.023
NNI(d)	0.020	0.051	0.113	0.013	0.023	0.050
GDI(d)	0.025	0.060	0.166	0.012	0.026	0.072

Table 2.8 and Figure 2.10 also confirm the ability of the WNNEM method to select a better control group in a harder situation. Notable differences can be seen in the cases of datasets 1, 3 and 5. However, it should also be noted that while the WNNEM method provided a more accurate control group for 17 datasets, in three cases the PSM did.

Figure 2.11 details the covariate imbalances separately. As can be seen, the WNNEM method in most cases was able to perfectly match on the covariate exhibiting a very high effect on treatment assignment (x_1), but the PSM could not. Also on the other covariates, the WNNEM method was able to achieve more balanced results.

My simulation results showed that the proposed WNNEM method achieved better results than the most widely applied from of PSM on both the benchmark dataset (Scenario I) and the more problematic datasets (Scenario II and Scenario III). As it was mentioned before, the WNNEM method considers the control group selection problem as a distance minimisation problem in the n-dimensional space. Distances

Table 2.8. Results of the Hansen and Bowers test in Scenario III.

		PSM			WNNEM	
dataset	$\|X_T\|$	$min(p)$	$avg(p)$	$max(p)$	p	diff.(p)
1	250	0.896	0.909	0.918	1.000	0.091
2	240	0.998	0.998	0.998	1.000	0.002
3	242	0.744	0.781	0.798	0.954	0.173
4	235	0.967	0.970	0.974	0.997	0.027
5	270	0.743	0.761	0.766	0.908	0.147
6	258	1.000	1.000	1.000	0.999	-0.001
7	269	0.998	1.000	1.000	0.967	-0.033
8	230	0.983	0.986	0.989	0.993	0.007
9	227	0.991	0.992	0.995	0.993	0.001
10	249	0.996	0.998	0.999	1.000	0.002
11	257	0.844	0.862	0.888	0.896	0.034
12	253	0.964	0.976	0.983	1.000	0.024
13	277	0.922	0.923	0.927	0.950	0.027
14	240	0.979	0.983	0.988	1.000	0.017
15	221	0.996	0.996	0.996	1.000	0.004
16	248	0.995	0.997	0.999	0.999	0.002
17	237	0.991	0.993	0.993	0.988	-0.005
18	252	0.973	0.975	0.976	0.996	0.021
19	256	0.989	0.989	0.989	0.998	0.009
20	255	0.996	0.997	0.998	1.000	0.003
min						-0.033
avg						0.028
max						0.173
sum						0.548

between the individuals of the case and control groups are calculated as weighted distances of the dimensions, and the aim is to match the control subjects to the case subjects in such a way, that the sum of their distances is minimal. It is easy to see that in a simple case, when the neighbour closest to an individual in the case group is chosen as the pair from the group of possible candidates, then minimisation problem is solved. The only problem arises when there are candidates which are closest to more than one individual of the case group. The WNNEM method solves the problem of conflicting subjects locally. In case of conflicting candidates, the WNNEM method also takes the second nearest neighbours into account, and the individual in conflict is matched to that individual in the case group for which the second nearest neighbour is farther away. However, solving the conflicts locally does not guarantee that the resulting control group is globally optimal. The

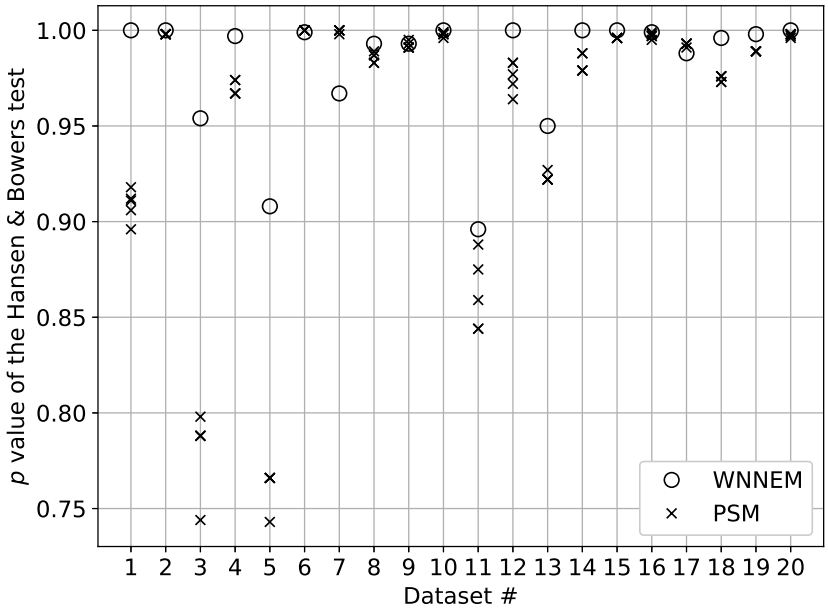

Fig 2.10. Results of the Hansen and Bowers test for each dataset in Scenario III.

Weighted Nearest Neighbour Control Group Selection with Simulated Annealing method to be presented in Section 2.3.3 aims to eliminate this problem.

2.3.3 Weighted Nearest Neighbour Control Group Selection with Simulated Annealing

In order to achieve a globally optimal solution, I developed a new control group selection method. The developed Weighted Nearest Neighbour Control Group Selection with Simulated Annealing (WNNSA) method [43] aims to achieve a globally optimal solution by applying simulated annealing (SA) [97]. The initial idea of the WNNSA algorithm comes from the field of optimisation.

Algorithms based on simulated annealing are such probabilistic algorithms that can find the global optima of a given function - however, it is not guaranteed. SA algorithms optimise the objective function (called energy, e) iteratively in the space of possible solutions such that they move the current state representing the actual solution of the problem into a new candidate state representing a new possible solution step-by-step. This moving is controlled by a probabilistic function which depends on the difference of the objective function of the current and neighbouring states, and a time-dependent variable called temperature (t). The main principle of the algorithm is that as time goes on (the temperature is decreasing), the probability that the algorithm will move to a state with higher energy (worse state) than the

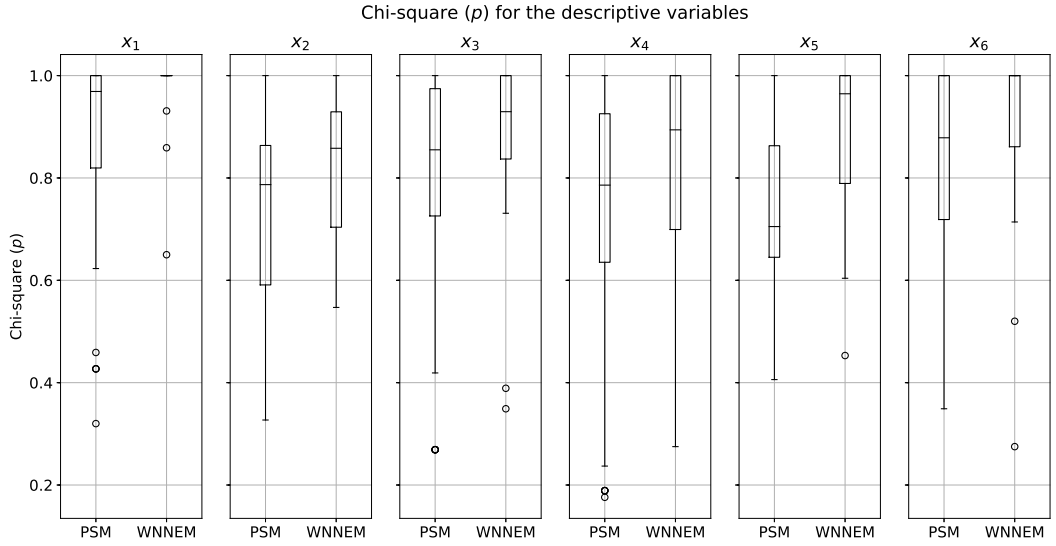

Fig 2.11. Distribution of all covariates in Scenario III.

energy of the current state decreases.

The WNNSA algorithm combines the simulated annealing approach with the distance calculation method applied in the WNNEM method. However, before we move on to the topic of simulated annealing, a shortcoming of the WNNEM method presented in Section 2.3.1, which appeared during the development of WNNSA, has to be addressed: the weight factors were only determined for covariates that positively affect the probability of the assignment to the treated group. To address this gap, the weight calculation method for the WNNEM, and as a result for WNNSA also, was extended the following way.

The calculation of the weights of covariates with OR value above 1 (positive association) does not change, and it can take any value above one as the weighting factor. In case of negative association (OR value in the range of $[0, 1)$), the weight of the covariate should be calculated as the reciprocal of the calculated OR value. In this way, the weights of the negatively associated covariates also take any value from $(1, \infty]$. Accordingly, the weights of covariates can be calculated as

$$w_i = \begin{cases} e^{b_i} & OR_i \geq 1 \\ \frac{1}{e^{b_i}} & OR_i < 1 \end{cases}. \tag{2.23}$$

Having the extended calculation of the weighting factors, the distance matrix containing the pairwise distances of the individuals in the treated and control groups (Eq. 2.18) can be calculated by weighting the dimensions. With the weight calculation fixed and clarified, we can move on to simulated annealing.

Each state in the search space represents a possible solution for the control group selection, meaning, each state represents a possible pairing of the individuals of the case group and the control group. The goal of the algorithm is to find the best pairing. To achieve this goal, the algorithm utilises the simulated annealing principle to select the best pairs for the treated individuals, and the goal is to minimise the sum of the pairwise distances of the paired individuals. The probability for selecting the candidate $\mathbf{X}_j \in X_C$ for the individual $\mathbf{X}_i \in X_T$ is calculated as

$$p(\mathbf{X}_i, \mathbf{X}_j) = \frac{p_{temp}(\mathbf{X}_i, \mathbf{X}_j)}{\sum_j p_{temp}(\mathbf{X}_i, \mathbf{X}_j)}, \tag{2.24}$$

where

$$p_{temp}(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{dist(\mathbf{X}_i, \mathbf{X}_j)^t} \tag{2.25}$$

and t is the temperature of the simulated annealing process.

The energy function (e) determining the fitness of the candidate solutions is given as

$$e = \sum_{(\mathbf{X}_i, \mathbf{X}_j) \in M} dist(\mathbf{X}_i, \mathbf{X}_j), \tag{2.26}$$

where $M = \{M_1, M_2, \ldots, M_m\}$ yields the pairing of the elements. In case of 1:1 matching, $m = |X_T|$. For later use, denote M_{i1} the first and M_{i2} the second element from the i-th pair from $M (i = 1, 2, \ldots, m)$.

In cases, when individuals to be paired can be selected from many candidates, many possible pairings are conceivable. To reduce the runtime of the algorithm, the WNNSA algorithm utilises linear cooling and looks for the optimal solution in a reduced search space. The applied heuristic constraints the search space such a way that an individual of the case group can only be paired to their k-closest neighbours from the candidate set. Further neighbours are not considered for pairing.

Denote $NN_k(\mathbf{X_i}, Y)$ the k-closest neighbours of $\mathbf{X_i}$ from the set Y. Using this notation, the k-size reduced environment for an individual $\mathbf{X}_i \in X_T$ is given by $NN_k(\mathbf{X}_i, X_C)$.

The WNNSA algorithm works relatively the same way as WNNEM, with some differences. Firstly, it is executed iteratively until the t temperature reaches 0. In each iteration, $p(\mathbf{X}_i, \mathbf{X}_j)$ is calculated for all elements of $NN_k(\mathbf{X}_i, X_C)$, then 1:1 matching with conflict resolution is performed. At the the end of each iteration, e is calculated for the current state, and is compared to the energy of the current best state (e_{best}). Finally, if the energy of the current state is lower than the energy of the

current best state, the current best state is replaced with the current state.

The detailed algorithm of the Weighted Nearest Neighbour Control Group Selection with Simulated Annealing method is presented in Algorithm 2.2. For the sake of clarity, it should be noted that $ActualMatching^{(t)}$ denotes a transient set of matched pairs that the algorithm generates at temperature t. $ActualMatching_i^{(t)}$ yields an element of this set, that is a specific matching of a case element with a candidate element. $ActualMatching_{i1}^{(t)}$ denotes the first and $ActualMatching_{i2}^{(t)}$ the second element of the matched pair of the i-th element from $ActualMatching^{(t)}$. The first element comes from the case group and the second one from the set of candidates to be paired as control individuals.

As Algorithm 2.2 shows, the WNNSA method uses a reduced environment for selecting the elements of the control group. However, the application of a reduced environment introduces another problem: below a given value of k, it is not guaranteed that the algorithm results in a control group with the desired size. This stems from conflicts occurring during the selection process. For example, consider the following situation for which a visual representation can be seen in Figure 2.12.

Let $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ be three individuals from the case group. Let $\mathbf{X}_4$ be the closest and $\mathbf{X}_5$ the second closest neighbour of $\mathbf{X}_1$ and $\mathbf{X}_2$ individuals among the candidate subjects. Furthermore, let $\mathbf{X}_5$ be the first and $\mathbf{X}_4$ the second nearest neighbour of $\mathbf{X}_3$. Moreover, let $\mathbf{X}_6$ be the third nearest neighbour of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$.

My aim was to select an equal-sized control group for the case group. In this case, if k is set to 2, then the reduced environments for $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ contain only the individuals $\mathbf{X}_4$ and $\mathbf{X}_5$. So, three paired control individuals cannot be selected from the reduced environments; therefore, 1:1 matching can not be performed.

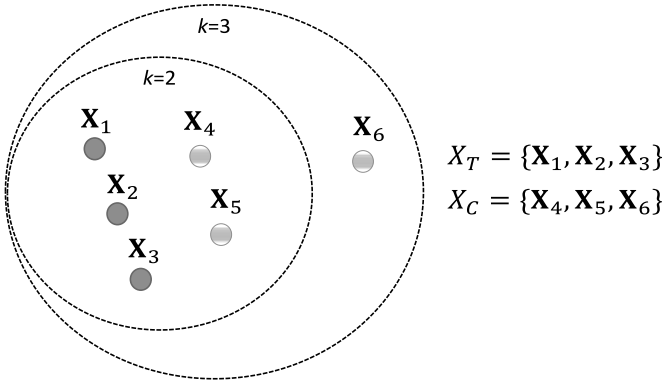

Fig 2.12. Demonstration of conflicting pairs in a reduced environment.

This problem can also be extended to higher k values. For this reason, a method to determine the minimal k value is needed. The problem of unsolvable conflicts

Algorithm 2.2: Weighted Nearest Neighbour Control Group Selection with Simulated Annealing (WNNSA)

Input: X_T: the set of the case group; X_C: the set of candidate individuals; k the size of the reduced environment; t_{max} the starting temperature

Output: X_{UT} the selected control group; M the set of the matched pairs

1 Initialise:
$X_{UT} = \emptyset$
$M = \emptyset$
$e_{best} = \infty$
$t = t_{max}$

2 Normalise X_T and X_C collectively using feature scaling.

3 Calculate the distance matrix **D** for all pairs of $\mathbf{X}_i \in X_T$ and $\mathbf{X}_j \in X_C$ by Eq. 2.18.

4 Determine $NN_k(\mathbf{X}_i, X_C)$ based on the distance matrix **D** for all $\mathbf{X}_i \in X_T$.

5 Determine $p(\mathbf{X}_i, \mathbf{X}_j)$ for each $\mathbf{X}_i \in X_T$ and for each $\mathbf{X}_j \in NN_k(\mathbf{X}_i, X_C)$ by Eq. 2.24.

6 Set:
$X^{(t)}_{unpaired} = X_T$
$X^{(t)}_{UT} = \emptyset$
$M^{(t)} = \emptyset$

7 Set $ActualMatchings^{(t)} = \emptyset$

8 For all $\mathbf{X}_i \in X^{(t)}_{unpaired}$
Select an $\mathbf{X}_j$ pair from $NN_k(\mathbf{X}_i, X_C)$ for $\mathbf{X}_i \in X^{(t)}_{unpaired}$ at random with probability $p(\mathbf{X}_i, \mathbf{X}_j)$.
Set $ActualMatchings^{(t)} = ActualMatchings^{(t)} \cup \{(\mathbf{X}_i, \mathbf{X}_j)\}$

9 For $l = 1, \ldots, |ActualMatchings^{(t)}|$
If $ActualMatchings^{(t)}_{l2}$ is selected for only one $\mathbf{X}_i \in X_T$
$X^{(t)}_{unpaired} = X^{(t)}_{unpaired} - \{\mathbf{X}_i\}$
$X^{(t)}_{UT} = X^{(t)}_{UT} \cup \{ActualMatchings^{(t)}_{l2}\}$
$M^{(t)} = M^{(t)} \cup \{ActualMatchings^{(t)}_{l}\}$

10 End if

11 Repeat *Steps 7 to 9*, till $X^{(t)}_{unpaired} \neq \emptyset$.

12 Calculate the actual energy $e^{(t)}$ for the matching $M^{(t)}$ by Eq. 2.26.

13 If $e^{(t)} < e_{best}$
$e_{best} = e^{(t)}$
$X_{UT} = X^{(t)}_{UT}$
$M = M^{(t)}$

14 Set $t = t - 1$.

15 Repeat *Steps 6 to 14* until $t = 0$.

described above can be solved mathematically.

As mentioned before, $NN_k(\mathbf{X}_i, Y)$ denotes the k-closest neighbours of $\mathbf{X}_i$ from the set Y. Additionally, $X^k_{C^*}$, the aggregated reduced set of candidates for the k-sized environment, can be calculated as

$$X^k_{C^*} = \{\mathbf{X}_j | \mathbf{X}_j \in NN_k(\mathbf{X}_i, X_C), \forall \mathbf{X}_i \in X_T\}. \tag{2.27}$$

Furthermore, denote those individuals from the case group for which $\mathbf{X}_j \in X^k_{C^*}$ is in their k-size reduced environment as $Dem(\mathbf{X}_j)$. $Dem(\mathbf{X}_j)$ is called the demand set of $\mathbf{X}_j$.

$$Dem(\mathbf{X}_j) = \{\mathbf{X}_i | \mathbf{X}_j \in NN_k(\mathbf{X}_i, X_C)\}, \tag{2.28}$$

where $\mathbf{X}_i \in X_T$.

Let $di(\mathbf{X}_j)$ be the demand index for $\mathbf{X}_j \in X^k_{C^*}$ quantifying those $\mathbf{X}_i \in X_T$ subjects which select $\mathbf{X}_j$ as one of the k-nearest neighbours into the k-reduced environment.

$$di(\mathbf{X}_j) = \frac{|Dem(\mathbf{X}_j)|}{k}, \tag{2.29}$$

where $|Dem(\mathbf{X}_j)|$ yields the size of the demand set of $\mathbf{X}_j$.

Denote the alternative selection index for $\mathbf{X}_j$ as $asi(\mathbf{X}_j)$, which quantifies the alternative selection options of $\mathbf{X}_j$ for all $\mathbf{X}_i \in Dem(\mathbf{X}_j)$. Alternative selection means that the elements of the demand set of $\mathbf{X}_j$ are paired to another candidate individual instead of $\mathbf{X}_j$.

$$asi(\mathbf{X}_j) = \frac{\sum_{\mathbf{X}_i \in Dem(\mathbf{X}_j)} min(di(NN_k(\mathbf{X}_i, X_C)))}{|Dem(\mathbf{X}_j)|}. \tag{2.30}$$

Using these metrics, the minimum size of the environment required for WNNSA to be successful can be easily defined: if exists such an $\mathbf{X}_j \in X^k_{C^*}$ that $di(\mathbf{X}_j) > 1$ and $asi(\mathbf{X}_j) > 1$, there is an unsolvable conflict. In this case, the size of the environment (that is the value of k) have to be increased. The method to determine the minimal value of k is summarised in Algorithm 2.3.

After the size of the minimal required reduced environment is determined, the WNNSA algorithm can be run. To perform a successful control group selection, the value of k must be set to at least the value determined by Algorithm 2.3. The higher the value of k is, the higher the degree of freedom the WNNSA algorithm has.

Algorithm 2.3: Determination of the minimal size for the reduced environment for the WNNSA algorithm

Input: X_T: the set of the case group; X_C: the set of candidate individuals
Output: k: the minimal size for the reduced environment

1 Calculate the distance matrix **D** by Eq. 2.18.
2 Set $k = 1$.
3 Determine $X^k_{C^*}$ by Eq. 2.27.
4 For all $\mathbf{X}_j \in X^k_{C^*}$:
  Calculate $di(\mathbf{X}_j)$ by Eq. 2.29.
  If $di(\mathbf{X}_j) > 1$
    Calculate $asi(\mathbf{X}_j)$ by Eq. 2.30.
    If $asi(\mathbf{X}_j) > 1$
      $k = k + 1$
      Go *Step 3*
5 Return k.

2.3.4 Evaluation of the proposed WNNSA method

To test the effectiveness of the extended WNNEM method and the WNNSA method, several Monte Carlo simulations were performed. In the following subsections, three scenarios are presented from them, which step by step show the effectiveness of the extensions introduced before. Scenario IV, which is based on Scenario I presented in Section 2.3.2.2, illustrates the applicability of the extension of the WNNEM method to negative covariates. In Scenario V, which utilises the same settings as Scenario I, the advantage of the WNNSA method using simulated annealing is presented. It is important to note, that in this scenario, negative covariates are not present. Finally, Scenario VI is a complex simulation containing both negative and positive covariates. This scenario aims to present the advantage of the WNNSA method in a rare feature space containing only a few covariates with few values.

In this research, the results of the extended WNNEM method and the WNNSA method were compared to two types of the PSM method and to stratified matching (SM) [98, 99]. The two types of the applied propensity score matching were the followings: (1) In practical studies, the PSM method is generally applied with a restrictive condition. This constraint is controlled by setting the *caliper size* parameter. Generally, the caliper size is set to 0.2 of the standard deviation of the logit of the propensity scores. It means that the control individuals can only be selected from a reduced environment of the treated elements. In the followings, this type of the PSM method is denoted as *PSM_02*. However, using this constraint, the control group selection method may also result in a control group that contains fewer indi-

viduals than the treated group. (2) In the second version of the PSM method, for a fair evaluation, the propensity score matching was run with dynamic caliper size. It means that the size of the neighbourhood (aka the caliper size) of the treated individuals was determined dynamically such that in each case, an appropriately sized control group could be selected. In the followings, this type of the PSM method is denoted as *PSM_DYN*. Such a fair evaluation was also used during the evaluation of the WNNEM method in Section 2.3.2. In the case of the WNNSA algorithm, the minimal size of the reduced k-size environment (k_{min}) was calculated in accordance with Algorithm 2.3. To increase the search space and the freedom of the algorithm, the value of k was set to $k = \lfloor k_{min} * 1.15 \rfloor$ in all scenarios.

As mentioned before, the effectiveness of the proposed methods was evaluated through Monte Carlo simulations. In each scenario, 100 independent datasets were generated with the given parameters. That is, each scenario was evaluated on 100 independent but similar datasets. As the WNNEM method is a deterministic algorithm, it was run only once on each generated dataset. In contrast, as the PSM_02, PSM_DYN, and WNSSA methods are not deterministic methods, they were executed 10-times on each dataset. For these methods, the best result from 10 runs was considered for the evaluation.

The quality of the selected control groups was evaluated from several perspectives. For distribution-based evaluation, the SMD, the t-test, the chi-squared test, the Hansen-Bowers test, and the Distribution Dissimilarity Index has been used. The pairwise similarities of the paired elements were evaluated by the Nearest Neighbour Index and by the Global Dissimilarity Index.

2.3.4.1 Datasets

Scenario IV is a modified version of the benchmark dataset used in Scenario I and presented in Section 2.3.2.2. I have modified it by changing the effect of some covariates of the dataset from positive to negative. As the original form of this dataset is a widely used dataset, I utilised it to show the efficiency of the simulated annealing in Scenario V. In Scenario VI, a novel, synthetic dataset was used for the simulations. In the following, these datasets are described in detail.

In Scenario IV, the logistic regression model to describe the probability for the treated group membership was formulated as described in [96], but the effect of the first, fourth and seventh covariates features was changed from positive to negative (Eq. 2.31). That means these features negatively affect the probability of belong-

ing to the treated group. The applied weight coefficients were same as in Section 2.3.2.2.

The settings of Scenario V are entirely in line with the settings of Scenario I presented in Section 2.3.2.2.

Scenario VI is a novel, synthetic dataset. This dataset contains fewer covariates than the previous two datasets, therefore, it better illustrates the problem of conflicting candidates. However, this dataset is more complex as it also contains covariates with negative and positive associations. Furthermore, it also contains nominal, ordinal and continuous variables. In this dataset, every individual is characterised by two ordinal variables with Binomial distribution ($x_j \sim B(4, 0.5), \quad j = 1, 2$), four binary variables with Bernoulli distribution ($x_j \sim B(0.5), \quad j = 3, \ldots, 6$) and two continuous variables with Normal distribution ($x_j \sim \mathcal{N}(2, 0.6), \quad j = 7, 8$).

$$\begin{aligned} \text{logit}(p_{i,treat}) = b_{0,treat} - \\ b_L x_{i1} + b_L x_{i2} + b_L x_{i3} - b_M x_{i4} + b_M x_{i5} +, \\ b_M x_{i6} + b_H x_{i7} - b_{VH} x_{i8} \end{aligned} \tag{2.31}$$

where $b_{0,treat} = -1.344090$, $b_L = \log(1.05)$, $b_M = \log(1.25)$, $b_H = \log(1.5)$ and $b_{VH} = \log(1.9)$. Approximately $19\,\%$ of the subjects were considered as members of the treated group.

2.3.4.2 Scenario IV - Illustration of the extended version of the WNNEM method

The example presented in Scenario IV illustrates the correctness of the modified weight calculation method given in Equation 2.23. For this purpose, the simulated datasets contain both positively and negatively affecting features.

Table 2.9 summarises the evaluations of the control groups selected by the SM, PSM_02, PSM_DYN, and the extended version of the WNNEM methods. Table 2.9 contains the minimal, average and maximal quality values of the control group selections performed on the generated 100 datasets. As DDI, NNI, and GDI metrics are distance measures, in their case, the lower the value, the more similar the selected control group is. In contrast, in the case of the Hansen and Bowers test (HB), the higher the value, the more similar the selected control group is. The maximum possible value of the HB test is 1.

It can be seen in Table 2.9 that the SM method resulted in the worst metrics. In no case was the method able to select a control group of the same size as the treated

Table 2.9. Quality measures for Scenario IV.

	SM			PSM_02			PSM_DYN			WNNEM		
	min	avg	max	min	avg	max	min	avg	max	min	avg	max
HB(p)	0.505	0.873	0.999	0.735	0.965	1.000	0.776	0.980	1.000	0.944	0.998	1.000
DDI(d)	0.535	0.603	0.704	0.023	0.053	0.123	0.008	0.019	0.034	0.006	0.013	0.021
NNI(d)	0.535	0.603	0.704	0.206	0.313	0.390	0.187	0.281	0.349	0.054	0.062	0.073
GDI(d)	0.535	0.603	0.704	0.234	0.364	0.475	0.214	0.364	0.475	0.057	0.068	0.084

group. Comparing the PSM_02 and the PSM_DYN methods, the advantage of the PSM_DYN method is clearly visible. When comparing the extended WNNEM and PSM_DYN methods, we can see that the extended WNNEM method could select more similar control groups than the PSM method with dynamic caliper size setting in more cases. This fact is confirmed by all four quality indicators. The results support that the proposed extension of the WNNEM method is appropriate for handling negative covariates.

As the Hansen and Bowers test is a widely used overall balance test, its values are presented in Figure 2.13 in detail for the 100 datasets. It can be observed that the interquartile range is the smallest in the case of the extended WNNEM method. Besides that, this method selects more similar control groups more often; it also works quite reliably.

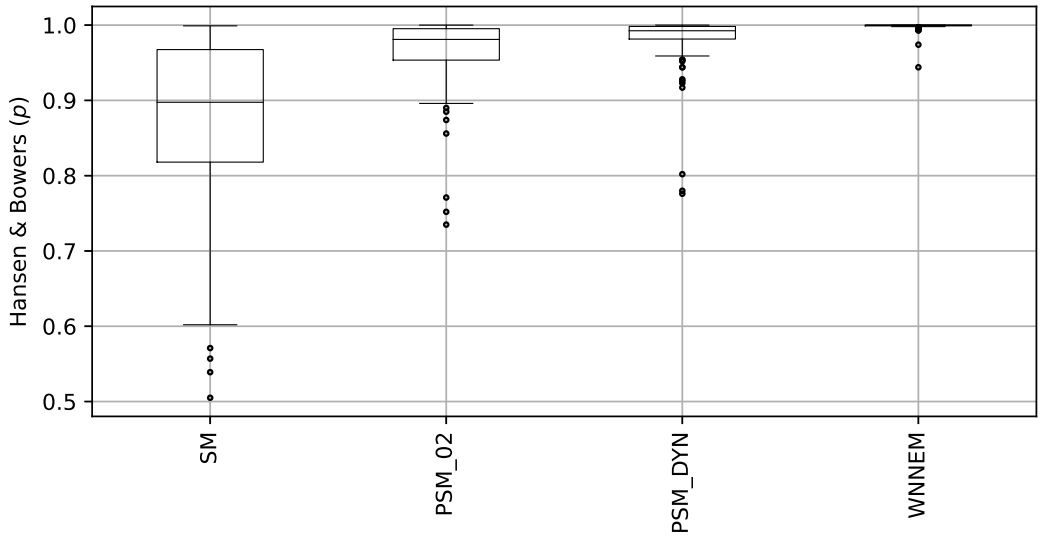

Fig 2.13. Variation of the Hansen and Bowers test in Scenario IV.

Figure 2.14 shows the individual balance values for the observed covariates separately. As the purpose of the present study was to examine how the weight-calculation of negative covariates works, this test is the most important test in this scenario. The similarity along with the covariates separately was calculated by the

Chi-square test, and the figure presents the distribution of the p values. Examining the properties separately, it can be seen that the PSM and WNNEM methods achieved better results than the stratified matching. The WNNEM method gave the best results for almost all variables, also including the negative covariates (x_1, x_4, x_7). Furthermore, SMD values were also calculated for all covariates and all matching methods. The SMD values for all matching were less than 0.1.

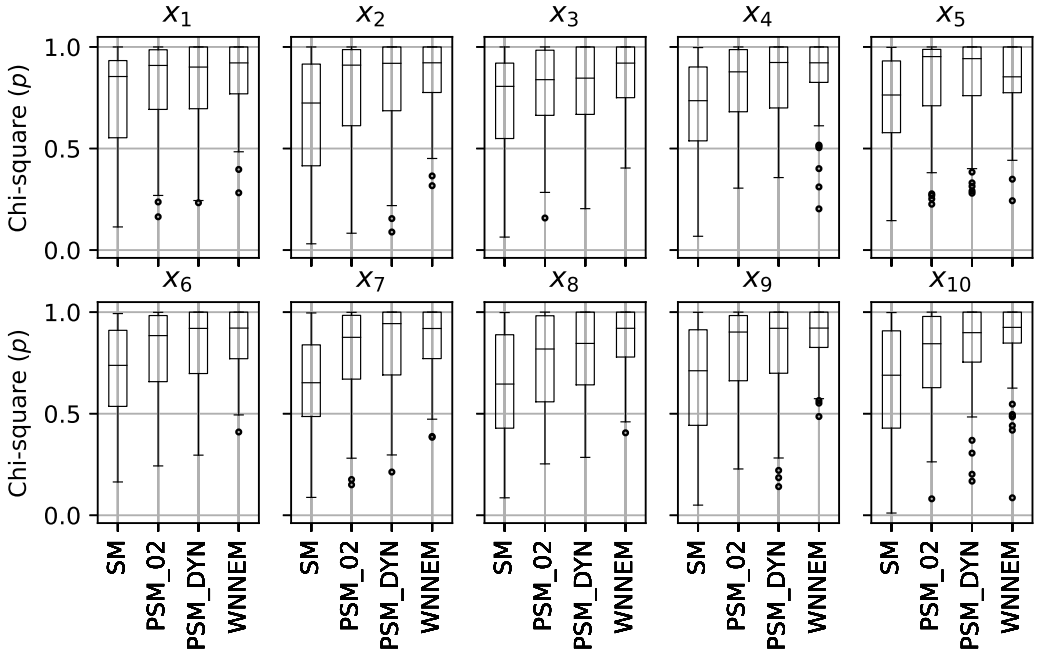

Fig 2.14. Distribution of all covariates in Scenario IV.

2.3.4.3 Scenario V - Illustration of the effectiveness of the WNNSA method against the deterministic WNNEM method

The example presented in Scenario V illustrates the advantage of the proposed WNNSA method against the WNNEM method. In this scenario, the data-generating process was identical to the one used in Section 2.3.2.2. The method of generating the datasets was not changed in order to illustrate the efficiency of the proposed method on a widely used benchmark dataset.

Table 2.10 presents the main quality indicators of the selected control groups. Considering the Hansen and Bowers test, the PSM, WNNEM, and WNNSA methods gave almost the same results. At the same time, the stratified matching resulted in less balanced control groups. The reason for the problem is again the same as before. As the dissimilarity indexes show, this method was not able to select full-sized

control groups. Besides the Hansen and Bowers test, the other distribution-based measurement (DDI) also confirms the similar qualities of the results of the PSM, WNNEM, and WNNSA methods. However, considering the neighbourhood-based indices (NNI, GDI), we can see that the WNNEM and WNNSA methods gave better results with one order of magnitude than the PSM methods.

Table 2.10. Quality measures for Scenario V.

	SM			PSM_02			PSM_DYN			WNNEM			WNNSA		
	min	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*
HB(p)	0.512	0.873	1.000	0.813	0.978	1.000	0.904	0.993	1.000	0.740	0.991	1.000	0.955	0.998	1.000
DDI(d)	0.504	0.574	0.631	0.021	0.061	0.116	0.006	0.014	0.023	0.006	0.012	0.022	0.005	0.011	0.021
NNI(d)	0.504	0.574	0.631	0.194	0.316	0.374	0.190	0.278	0.325	0.052	0.060	0.070	0.056	0.070	0.080
GDI(d)	0.504	0.574	0.631	0.214	0.348	0.416	0.212	0.313	0.367	0.052	0.061	0.077	0.062	0.073	0.097

If we compare the WNNEM and WNNSA methods (Table 2.10), we can see that in terms of neighbourhood indices, the WNNSA method performed slightly worse than the WNNEM method. The reason for this is that WNNSA does not always select the nearest neighbours. In contrast, as the WNNSA is trying to achieve a globally optimal solution, this method gave better results in terms of the indices measuring the distributions of the whole dataset (Hansen and Bowers test, Distribution Dissimilarity Index). In consequence, the variable-wise balance may be a little bit more diverse in the case of the WNNSA method (Figure 2.15). However, the SMD values for all matching methods were less than 0.1.

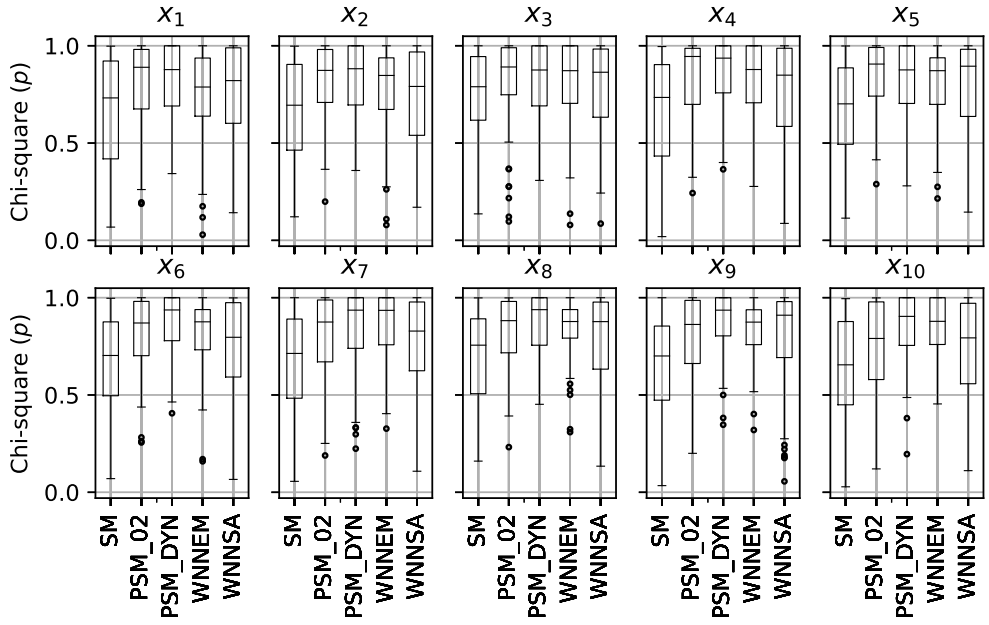

Fig 2.15. Distribution of all covariates in Scenario V.

At the same time, Figure 2.16 shows that the interquartile range of the WNNSA

method is smaller than the interquartile range of the extended WNNEM method. That is, the WNNSA method can select more similar control groups more reliably. For better visibility, Figure 2.16 does not include the results of the SM method as its outlier values were too low. For the sake of completeness, the first quartile ($Q1$) of data for SM is equal to 0.8092, the median of the data ($Q2$) is equal to 0.9235, and the third quartile of data ($Q3$) is equal to 0.9730.

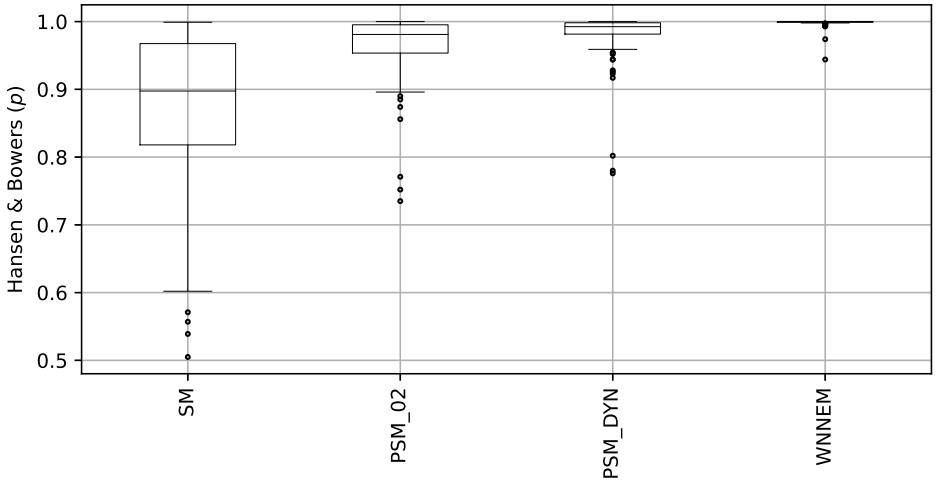

Fig 2.16. Variation of the Hansen and Bowers test in Scenario V.

2.3.4.4 Scenario VI - Advantage of the WNNSA method in a more conflicted environment

The results presented in Sections 2.3.4.2 and 2.3.4.3 were based on benchmark datasets. As the proposed WNNSA method aims to improve the efficiency of the WNNEM method in a conflicted environment, the main advantage of the presented method can be primarily presented with such a kind of dataset.

Table 2.11 shows the values of different quality measures of the control groups selected by the SM, PSM_02, PSM_DYN, the extended version of the WNNEM, and WNNSA methods in Scenario VI. Table 2.11 contains the minimal, average and maximal values for the generated 100 datasets.

Table 2.11. Quality measures for Scenario VI.

	SM			PSM02			PSMDYN			WNNEM			WNNSA		
	min	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*
HB(p)	0.140	0.724	0.995	0.523	0.941	1.000	0.729	0.960	1.000	0.769	0.969	1.000	0.815	0.991	1.000
DDI(d)	0.617	0.710	0.800	0.062	0.102	0.162	0.050	0.072	0.102	0.032	0.056	0.078	0.034	0.052	0.071
NNI(d)	0.718	0.789	0.858	0.591	0.661	0.705	0.528	0.640	0.678	0.285	0.303	0.321	0.300	0.318	0.335
GDI(d)	0.637	0.728	0.815	0.314	0.411	0.463	0.279	0.376	0.446	0.035	0.046	0.057	0.043	0.056	0.069

It can be seen in Table 2.11 that the SM method yielded the worst results in most cases, analogously to Scenario IV and Scenario V. Overall, the nearest neighbour-based methods (WNNEM and WNNSA) achieved better results than the PS-based methods (PSM_02 and PSM_DYN). The differences between the two groups are similar in magnitude as in Scenario V.

In terms of NNI and GDI measurements, the deterministic WNNEM method achieved better results than the non-deterministic WNNSA method. However, it can be seen that in terms of overall balance (HB(p)) WNNSA achieved better results. The differences between the two methods are greater in this case than in Scenario V. This fact is also observable in Figure 2.17. In this figure, the results of the SM method are again not presented. For the SM method, the values are the followings: $Q1 = 0.6105$, $Q2 = 0.7810$, and $Q3 = 0.9093$.

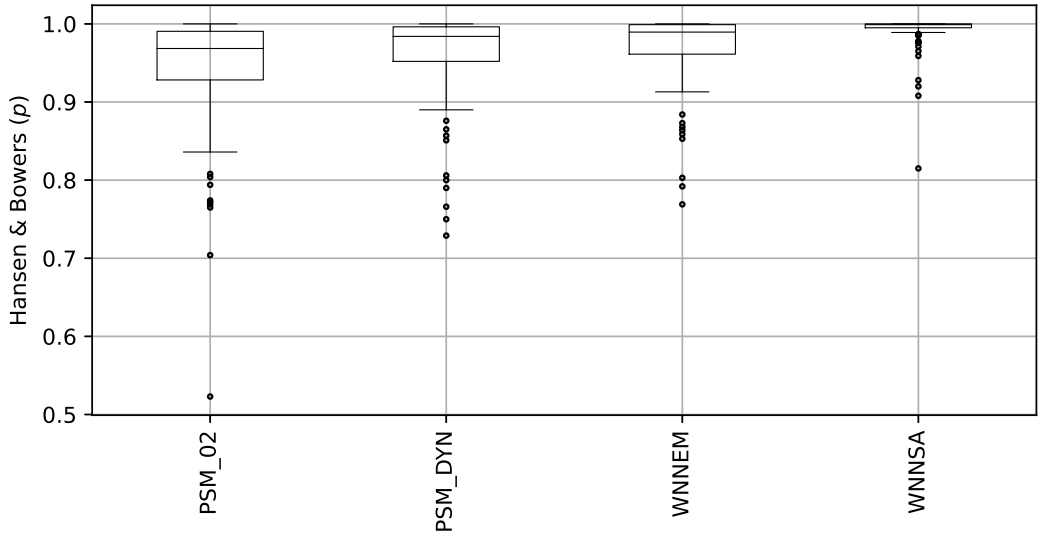

Fig 2.17. Variation of the Hansen and Bowers test in Scenario VI.

Figures 2.18 and 2.19 show the individual balance values for continuous and non-continuous variables. In the case of non-continuous variables (Figure 2.18), the WNNSA method achieved the best results in all cases. Comparing the WNNSA method to the WNNEM method, the distribution of the balance in the case of x_1 and x_2 variables is better in the case of the WNNSA method; in the case of the other covariates, it is the same. In the case of continuous variables (Figure 2.19), the WNNEM method gave less good results than the PS-based methods, but the results of the WNNSA method are similar and better for x_7. The SMD for all matching methods were again less than 0.1.

To sum up, the WNNSA method can be seen as an improvement of the Weighted

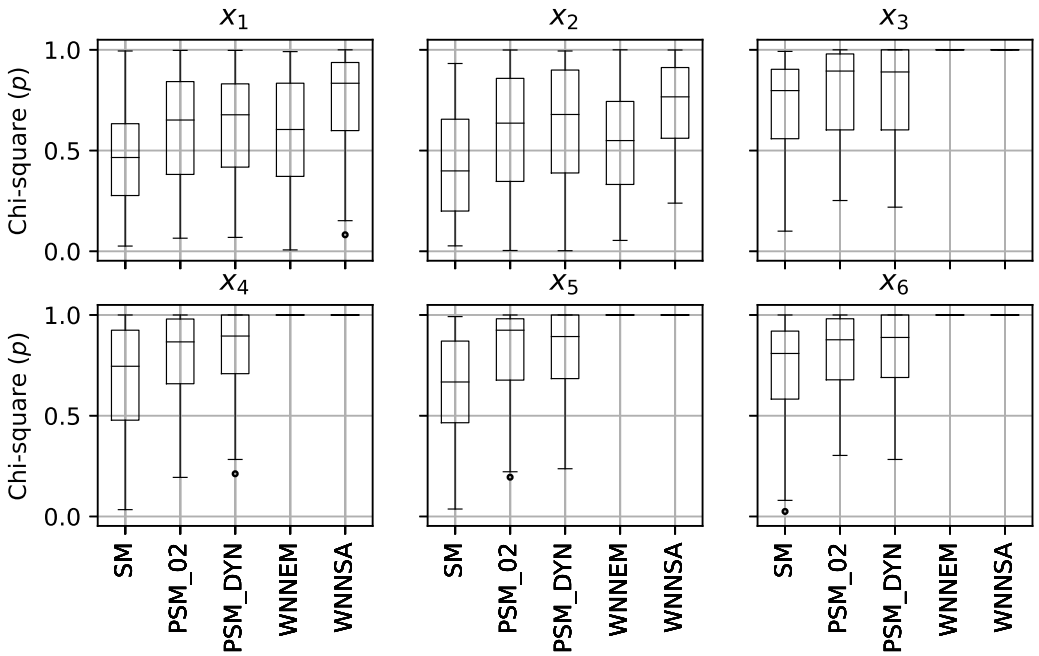

Fig 2.18. Distribution of nominal and ordinal covariates in Scenario VI.

Nearest Neighbour Control Group Selection with Error Minimization (WNNEM) method. It utilises simulated annealing to achieve a global optimal solution and to find the best pairing of individuals from the case group and the control group. All optimisation decisions are made in a reduced environment. My results showed that the WNNSA method can achieve better results than the WNNEM method and also has its advantage against the propensity score matching methods in rare feature spaces.

The last question that I wanted to answer during my research regarding control group selection was how missing variables effect the outcome of case-control studies? The details and results of my research can be found in the next section.

2.4 Measuring the effect of missing variables

As retrospective cohort studies look back in time, they do not require a long time for collecting data about patients. However, these studies must face the fact, that the range of available data is not always complete. For this reason, it may happen that case and control groups differ not only in the previously planned characteristic property (e.g., medication treatment vs placebo), but hidden differences may exist as well, for which we do not have data. Of course, similar cases may also occur in

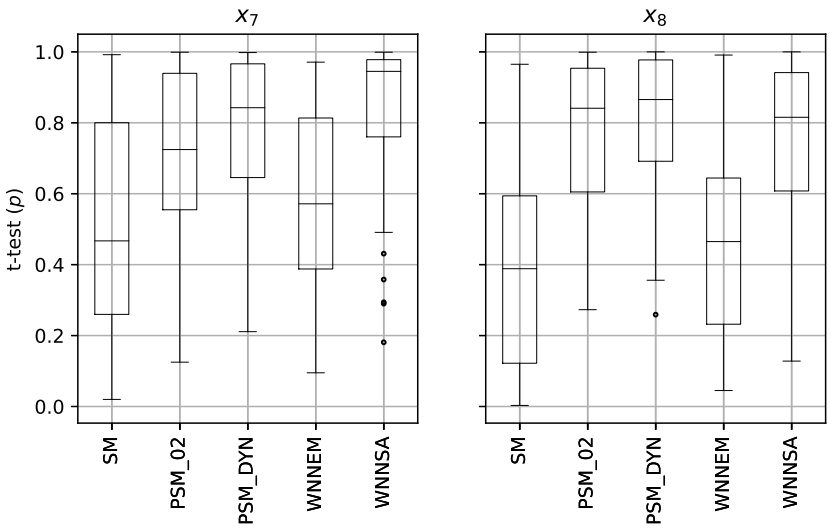

Fig 2.19. Distribution of continuous covariates in Scenario VI.

prospective studies, if the scope of data included in the study is not complete.

The effect of the known independent variables on the dependent variable (outcome variable) can be determined in different ways. If the outcome variable (e.g., the appearance of a disease) is categorical, logistic regression is the most commonly used method for this purpose. However, LR is based on estimation and it includes uncertainty. The uncertainty of the model can be measured by the R^2 value, which suggests how well the observed outcome is replicated by the model from the independent variables. If the established logistic regression model is inaccurate, then the value of the output variable (e.g., the appearance of a disease) cannot be predicted with sufficient certainty.

However, the question arises, whether the uncertainty of the model can be derived from the missing variables. Furthermore, the uncertainty of the prediction of the output variable only arises from the predictive variables included in the model or the effect of the missing variables may also affect this uncertainty? How does the uncertainty of the model relate to the uncertainty of the prediction of the output variable?

During my research, I used a statistics-based approach with which I tried to determine the relationship between the accuracy of the binary logistic regression model and the uncertainty of the prediction of the output variable. The analysis was based on benchmark datasets generated with Monte Carlo simulations. Using these datasets, binary logistic regression-based propensity score matching was performed under various conditions to generate possible control groups, and then the deviation of the output variable in the case group and the control group was investigated in

order to determine the degree of distortion.

2.4.1 The methodology of the research

The effect of the known independent variables on the outcome variable is expressed by the calculations of odds ratios using logistic regression. As logistic regression is a probabilistic model that does not guarantee that the regressed outcome is entirely describable with the independent variables, it is possible to measure this uncertainty and there are various methods to do so.

The most basic measure is the coefficient of determination, denoted by R^2. R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variables. It provides a measure of how well a model approximates the observed outcomes based on the proportion of total variation of outcomes explained by the same model. Usually, the value of R^2 is in the range of $[0, 1]$. The better the linear regression fits the data, the closer the value of R^2 is to 1. However, values of R^2 outside the range of $[0, 1]$ can occur, depending on the used measure.

R^2 does not indicate whether the independent variables cause the changes of the dependent variable or omitted-variable bias exists. There is no way to tell if the correct regression was used, if the most appropriate set of independent variables has been chosen or if there is a collinearity present in the data on the explanatory variables. The model might be improved by using transformed versions of the existing set of independent variables, and it is possible that there are not enough data points to make a solid conclusion. It is important to take note of the second caveat: R^2 does not indicate whether omitted-variable bias exists. But still, R^2 provides a measure to quantify the model quality.

My main research aim was to discover if there is a measurable numeric relationship or determined correlation between the value of the general R^2 of models with omitted independent variables and the influence of the omitted independent variable on the outcome variable. My research was based on the assumption that, if the set of observed variables is complete, the logistic regression model properly describes the relationship between the independent variables, and the dependent variable and the R^2 value of the model is around 1.0. Selecting a control group to a sample based on this model guarantees that the deviation of the outcome variable is marginal between the case group and the control group based on the assumption that the odds ratio values are adequate.

The effect of the missing variables was analysed the following way. Benchmark

datasets were generated by Monte Carlo Simulation. The investigation scenario consisted of 100 simulated datasets of a 1000 individuals characterised by 8 binary independent variables. All 8 independent variables $(x_1, \ldots, x_8)$ were independent Bernoulli random variables with a probability parameter of 0.5. These independent variables model the characteristics of a certain patient, e.g., sex, diagnoses and other descriptors.

For each dataset, additional datasets were created: for every independent variable, only one was omitted and the others were kept intact. This resulted in 8 additional datasets, each one containing only 7 independent variables. This way the number of investigated datasets totalled 900 $((1 + 8) * 100)$.

To determine the output variable, a utility value $(y^{'})$ was calculated for each individual based on Eq. 2.32. It can be seen that the values of the regression coefficients were chosen in such a way, that the effect of the independent variables changes uniformly from 1.0 to 3.0. Therefore, omitting x_1 should have a lower effect on the R^2 value of the model than omitting x_8.

$$y^{'} = 1.0x_1 + 1.2x_2 + 1.4x_3 + 1.6x_4 + 1.8x_5 + 2.0x_6 + 2.5x_7 + 3.0x_8. \quad (2.32)$$

The binary outcome (y) was determined individually for each dataset.

$$y = \begin{cases} 1 & \text{if} \quad y^{'} > median(y) \\ 0 & \text{otherwise} \end{cases}, \quad (2.33)$$

where $median(y)$ is the median of all y values. In a more comprehensible way, if the exposure of an element from a specific dataset was higher than the median of all elements from the same dataset, the outcome is 1 (having a diagnosis or receiving a treatment), otherwise 0. This way the probability of the outcome estimates 0.5 for each specific dataset.

After the creation of the datasets, to estimate the propensity scores of the individuals and the R^2 values for the models logistic regression was performed on each of them independently. In the next step, I determined the R^2 difference values for each coherent dataset as

$$d_{R^2 i} = abs(R^2_{baseline} - R^2_{x_i}), \quad i \in 1, \ldots, 8, \quad (2.34)$$

where $R^2_{baseline}$ is the R^2 value of the dataset containing all independent variables and $R^2_{x_i}$ is the R^2 value of the dataset from which x_i was omitted.

In the next step, test groups were created by blind random selection from each dataset, having the outcome retain 0.5 probability. The remaining elements formed the population, which contained the possible entities of the control groups. 50 control groups were selected for each test group with propensity score matching (caliper size=0.05). The evaluation of the results was based on the average values of the 50 control groups. The individuals of the control groups were selected in two different ways.

In the first case, called *realistic case*, I assumed, that the population giving the basis of the control group contains individuals both with 1 and 0 values of the omitted binary variable. This case simulates when the population from which the control group is selected may contain random values on the missing predictive variable.

In the second case, called *pessimistic case*, the worst case was modelled, when the population contains only such individuals where the value of the invisible predictive variable was equal to 1. This is the case, when we do not know, for example, that diabetes has a great impact on the outcome variable, and we select people into the control group without taking into consideration this feature, and the resulted control group contains only diabetic patients.

As the output variable in my investigation was binary, the distribution of the output was determined as the ratio of cases with $y = 1$ value, which models, for example, the frequency of a disease. So, the relative difference of the output variable in the case group and the control group was calculated as

$$rel_{err} = \frac{|\{\mathbf{X}_i \in X_{UT} \mid y_i = 1\}|}{|\{\mathbf{X}_j \in X_T \mid y_j = 1\}|}. \tag{2.35}$$

Finally, I compared the calculated $d_{R^2 i}$ and rel_{err} values.

2.4.2 Findings of the investigation

Figure 2.20 shows the relationship between the probability of the outcome being 1 and the $d_{R^2 i}$ value. The left side of Figure 2.20 shows that there is no noticeable relationship between the accuracy of the logistic regression model and the probability of the outcome in the realistic scenario. The quality of the selected control groups is almost the same in every case. The deviation of the probability of the outcome from the expected value (shown in the figure as a cyan horizontal region which represents the minimum, average and maximum probability of the outcome being 1 calculated based on the case groups) is within a $10\,\%$ range. It seems that

the quality of the outcome variable is not affected by the quality of the model. The right side of Figure 2.20 (pessimistic scenario) shows a more noticeable connection. The omitted variable strongly affects the value of the outcome variable. The worse the logistic regression model estimates the outcome, the bigger the difference is in the probability of the outcome variable. The more inaccurate the model, the higher the probability of the outcome being 1.

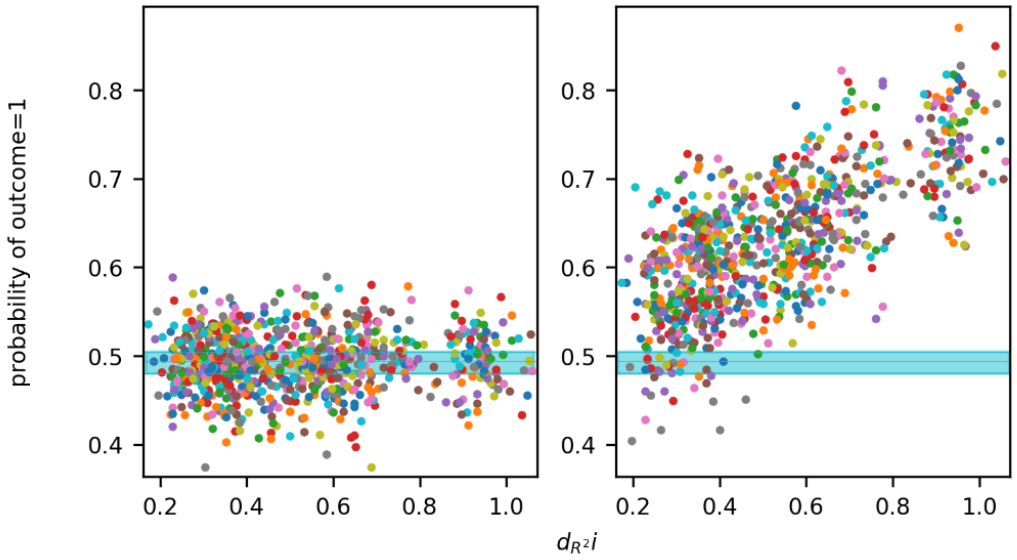

Fig 2.20. Relationship between the probability of the outcome being 1 and the d_{R^2i} values for the realistic scenario (left) and pessimistic scenario (right).

Figure 2.21 shows the relative error of the probability of the outcome being 1 between the case and the selected control groups as a function of d_{R^2i}. Just as previously, on the left side (realistic scenario) there is no noticeable relationship and the relative error tops at $20\,\%$. In contrast, in the pessimistic scenario (right side) there is a linear relationship between the measures. The higher the inaccuracy of the model, the higher the relative error becomes, reaching even $70\,\%$.

The results of the logistic regression based analysis are influenced by the ignoration of an explanatory binary variable. In the realistic case, when the omitted explanatory variable can take any value in the control group, the relative error of the predicted dichotomous value moves between $0\,\%$ and $20\,\%$. However, if the omitted explanatory variable only takes 1 as value in the control group, the relative difference between the predicted dichotomous outcome value with an omitted explanatory variable and the outcome value without any omitted explanatory variable can reach $70\,\%$.

To sum up, we can see that the selection of independent variables is a critical

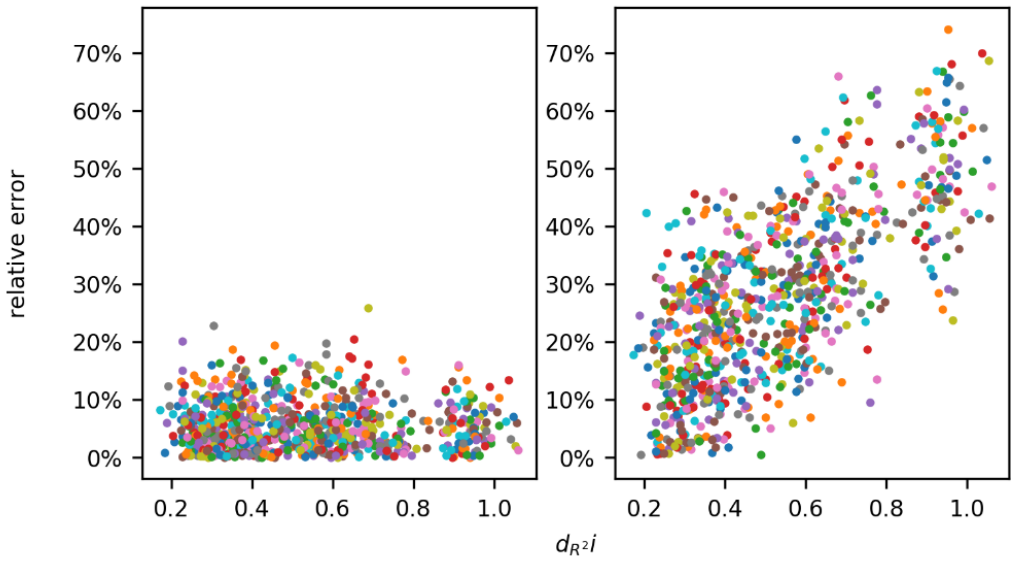

Fig 2.21. Relative error of the probability of the outcome being 1 as a function of $d_{R^2i}$for the realistic scenario (left) and pessimistic scenario (right).

step in case-control studies. The results of case-control studies rest on a correctly constructed dataset and adequate control group selection. Missing variables that have a high effect on the outcome variable may significantly distort the analysis results, so during the design phase determining such variables and enrolling them into the study is an essential task.

2.5 Related theses

Thesis 1.1

I proposed three quantitative dissimilarity measures to measure the dissimilarity of case and control groups regardless of the types of variables. Two of them evaluate the similarities of case and control groups based on the similarities of the paired individuals, and the third one compares the distributions of the characteristic features of the groups. The characteristics of the proposed methods was shown on synthetic datasets. All proposed measures are linear but their responsiveness is different. Results pointed out the fact that evaluating case and control groups must be made from different aspects, using both pairwise and distribution-based measures.

Thesis 1.2

I proposed a novel nearest neighbour-based control group selection method called Weighted Nearest Neighbours Control Group Selection with Error Minimization (WNNEM). The proposed method calculates the dissimilarities of the individuals in the original feature space of the independent variables. The independent variables are weighted based on a logistic regression-fit. For finding the nearest neighbours, WNNEM uses Vogel's approximation to solve such cases where an individual of the candidate group would be paired to more than one individual of the case group. The effectiveness of the WNNEM method was evaluated on benchmark and synthetic datasets. Evaluation results showed that the proposed WNNEM method is able to select a more balanced control group than the most widely applied greedy form of the propensity score matching method.

Thesis 1.3

As the previously developed WNNEM method utilises local optimisation, I proposed a novel simulated annealing-based control group selection method called Weighted Nearest Neighbour Control Group Selection with Simulated Annealing (WNNSA). The WNNSA method utilises simulated annealing to achieve a global optimum during control group selection to find the nearest neighbours. The effectiveness of the WNNSA method was evaluated on benchmark and synthetic datasets. Evaluation results showed that the proposed WNNSA method is able to select a more balanced control group than the WNNEM method if numerous conflicted situations arise in the selection process of similar individuals.

Thesis 1.4

I analysed the effect of missing binary independent variables on the results of case-control studies using logistic regression-fit. Using Monte Carlo simulations, my empirical results showed that there is a correlation between missing binary independent variables and the model accuracy. The Monte Carlo simulations revealed, that the selection of independent variables is a critical step in case-control studies as a biased control group regarding the missing variable may crucially affect the analyses results.

Related publications

P1 Szabolcs Szekér and Ágnes Fogarassyné Vathy. Kontrollcsoport generálási lehetőségek retrospektív egészségügyi vizsgálatokhoz. *Orvosi Informatika 2016 A XXIX.Neumann Kollokvium konferenciakiadványa*, Neumann János Számítógép-tudományi Társaság, pages 135-139, 2016.

P2 Szabolcs Szekér, György Fogarassy, and Ágnes Vathy-Fogarassy. Comparison of control group generating methods. *Studies in Health Technology and Informatics*, Vol. 236, pages 311-318, 2017. (Q3)

P3 Szekér Szabolcs, Fogarassyné Vathy Ágnes. Látens változók hatása dichotom kimenetű vizsgálatok kiértékelésére. *Orvosi Informatika 2018 A XXXI. Neumann Kollokvium konferencia-kiadványa*, Neumann János Számítógép-tudományi Társaság, pages 37-42, 2018.

P4 Szabolcs Szekér and Ágnes Vathy-Fogarassy. The effect of latent binary variables on the uncertainty of the prediction of a dichotomous outcome using logistic regression based propensity score matching. *Studies in Health Technology and Informatics*, Vol. 248, pages 1-8, 2018. (Q3) *Best PhD Paper Award*

P5 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Measuring the similarity of two cohorts in the n-dimensional space. *The 11th Conference of PhD Students in Computer Science: Volume of short papers CS2*, pages 151-154, 2018.

P6 Szekér Szabolcs, Fogarassyné Vathy Ágnes. Kontrollcsoport kiválasztása súlyozott k-nn módszer alkalmazásával. *Orvosi informatika A XXXII. Neumann Kollokvium konferencia-kiadványa*, Neumann János Számítógép-tudományi Társaság, pages 7-12, 2019.

P7 Szabolcs Szekér and Ágnes Vathy-Fogarassy. How can the similarity of the case and control groups be measured in case-control studies? *Proceedings of IEEE International Work Conference on Bioinspired Intelligence IWOBI 2019*, IEEE, pages 33-40, 2019.

P8 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Weighted nearest neighbours-based control group selection method for observational studies. *Plos One*, 15(7): e0236531, 2020. (D1, IF: 3.24)

P9 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Optimized weighted nearest neighbours matching algorithm for control group selection. *Algorithms*, 14(12): 356, 2021. (Q2)

Related abstracts

A1 Szekér Szabolcs, Ágnes Vathy-Fogarassy. Novel k Nearest Neighbour-based Control Group Selection Methods. *13th Miklós Iványi International PhD & DLA Symposium - Abstract Book: Architectural, Engineering and Information Sciences*, Pollack Press, page 124, 2017.

Chapter 3

Information extraction from echocardiography documents

Extracting information from free-text written medical texts is a challenging task and has been the subject of numerous publications in recent years [100, 101, 102, 103, 104, 105]. The difficulty of the task stems from the fact that free-text written medical documents follow less the grammatical rules of the written language, contain many abbreviations and spelling mistakes, and the terminology used is typically arising from several languages.

Processing Electric Medical Records written in free-text requires different Natural Language Processing (NLP) techniques. The information extraction process generally includes two main steps: named-entity recognition (NER) and relation extraction (RE) [50]. NER aims to identify names or entities (e.g., diseases, medical tests, results of tests), while RE aims to identify relations between them (e.g., symptoms related to diseases).

NER is usually implemented using direct search, rule-based search, machine learning methods or their combinations [106]. In practice, searching methods using regular expressions defined by medical experts are most commonly used (e.g., [107, 108]). The drawback of this approach is that it is difficult to provide a sufficiently general yet specific, complex regular expression that can handle typos and different wordings. Furthermore, they are mainly developed for extracting only one or some predefined keywords. Rule-based methods (e.g., [109, 110]) apply rules defined by experts during the search. Machine learning methods (e.g., [111, 112]) also show good performance in recognizing entities; however, their performance is highly influenced by the corpus used for training the model [113]. More recently,

for exploring the text descriptions more effectively, text mining, NLP-based (e.g., [114, 115, 116]), and neural network-based (e.g., [117, 118]) methods have been developed, and their use is becoming more widespread. From deep learning methods, mainly the recurrent neural networks (RNN), long short-term memory networks (LSTM), its bidirectional version (Bi-LSTM), the pre-trained transformers models (e.g., Bidirectional Encoder Representations from Transformers, BERT) and convolutional networks (CNN) are used [119, 120]. The authors of these articles generally point out that these models require a considerable computational capacity to build, and the literature review shows that these models perform well mainly in the field of Chinese medical NER due to the specific structure of Chinese written medical text records [121]. But there are also examples of their use in other languages for named-entity recognition, such as in Spanish or Swedish clinical texts [122].

Considering the application areas, we can find many biomedical application examples, but information extraction from cardiac ultrasound findings is one of the most frequently researched areas. Since the ejection fraction (EF) is one of the most important diagnostic measures and one of the strongest prognostic indicators in patients with cardiovascular disease, several studies aim to extract this diagnostic result (e.g., [52, 53, 54, 123]). Garvin et al. [53] used regular expressions for extracting EF values from echocardiogram reports. A similar method is also found in [58]. In [54], the authors also aimed to extract EF information from three kinds of clinical documents. Based on the characteristics of the corpora, rule-based regular expressions and machine learning-based NLP methods were applied. The authors evaluated the methods from different aspects, including the quantitative evaluations of the extraction of EF mentions, mentions of left ventricular systolic function (LVSF), extraction of EF quantitative values, and EF or LVSF qualitative assessments. Naturally, in addition to the extraction of the EF values, other research was also published that aimed to obtain other cardiac ultrasound characteristics. For example, Wells et al. [56] utilized NLP-based parsing and outlier analysis to extract flow velocities and chamber dimensions.

In addition to the extraction of predefined characteristics, some studies aimed to extract a broader range of information. In [55], pattern matching was applied to identify the relevant terms, and a concept-mapping algorithm was developed to assign the terms to the appropriate measurement concept. Kaspar et al. [124] investigated how all variables could be extracted from echocardiogram reports and what their quality would be for secondary use. The main conclusion of their study was that data could be extracted from echocardiography documents, but extraction

processes should be treated with caution, as the time and effort spent defining every variable may make it dubious.

Based on previous publications, we can see that most of the solutions developed aim at extracting only one or a few predefined results, typically using pattern matching or integrating corpus-specific knowledge. Only two studies [55, 124] have been published on comprehensive measurement outcome extraction, but they integrate a big amount of corpus-specific knowledge.

To overcome these shortcomings, I developed a corpus-independent method to extract quantitative measurement results from echocardiography documents. As the proposed method utilises text-similarity-based mapping, I have analysed different text-similarity measures to find the most suitable measure for information extraction from echocardiography documents. The developed method automatically identifies the name of the measurements and their recorded results in the text, and returns them in a structured way. The efficiency of the method has been evaluated and presented on a large corpus of Hungarian echocardiography documents.

The rest of this chapter is organised as follows: Section 3.1 introduces the corpus used during my research and presents the challenges of extracting information from echocardiography documents. In Section 3.2, different text-similarity measures are evaluated to find the most suitable measure for the developed method. Finally, Section 3.3 discusses the developed method in detail and evaluates it using the corpus introduced in Section 3.1.

3.1 Corpus

Echocardiogram is a sonogram of the heart. It is one of the most widely applied diagnostics test in cardiology: routinely used in diagnosis, management and follow-up of patients with any suspected or known heart disease. Echocardiography reports usually contain two parts in terms as diagnostic content: a semi-structured part where results are usually stored in term-value pairs (e.g., EN: "Septum: 14 mm", HU: "Szeptum: 14 mm") and a free text part written in natural language (e.g., EN: Left "ventricular hypertrophy" HU: "Koncentrikus bal kamra hypertrophia"). The form and content of the reports differ in medical institutes. The form and the content of the reports are mainly determined by the habits of the medical assistants and doctors. For example, the test result is separated by a colon from the test name in

some sites, while others do not use any separator character. Space characters typically separate different measurement results from each other, but other separator characters, such as semicolons, can also be found. Furthermore, even within one site, the same test result can be recorded with or without a unit of measurement in the reports. There is also a wide variation in how missing data are marked; furthermore, typographical errors increase the variety of the documents. A translated example of an echocardiography document can be seen in Figure 3.1.

Ao. root: 38 mm
Left atrium M-mode 45 mm.
Septum end diastolic 12 mm syst 14 mm Posterior wall end diasolic: 10mm syst 15mm left ventr diast 55 mm, syst 30mm
2D right ventricular diast basal: 40 mm.
ejection fraction (visual estimation) 60 %

Mildly dilated left atrium. Mild concentric left ventricular hypertrophy. Aortic valve is sclerotic, but the opening is normal. Moderate to severe aortic regurgitation. Mild calcified mitral apparatus, mitral valve regurg. gr. I.
Trace tricuspid regurgitation. Pulmonic valve regurgitation: I-II. No obvious wall motion abnormalities.

Fig 3.1. Raw echocardiography report translated to English.

The effectiveness of my proposed method was evaluated by processing 20,074 echocardiography reports. The document set was collected in a Hungarian hospital and contained all findings recorded between 2017 and 2021 by multiple physicians. The findings were anonymised, and did not contain any information about the patient or the examining physician. As there is no publicly available benchmark dataset to evaluate such methods, a dataset had to be collected, and, as none of the proposed methodologies contain any language-dependent processing activity and the proposed method can be used for documents written in any language, the collected dataset was adequate for evaluation.

3.1.1 Challenges of processing echocardiography documents

The methods presented in the introduction of Chapter 3 are widely applicable to extract information from medical documents mainly written in English, however the nature of Hungarian language requires specific tools to extract information from

medical documents written in Hungarian. In this section, I discuss the How To-s and challenges of the general extraction process of echocardiography reports, and also present some Hungarian language specific problems.

As I mentioned before, the semi-structured part of echocardiography reports contains medical information in term–value pairs separated by colon. The term part refers to the identifiable named entities and the value part refers to the measured and recorded value for that named entity. The measured value may also contain a unit of measurement. However, based on the extraction approach, various challenges emerge aside from typographical errors during term extraction. These challenges are described in detail in the following subsections.

Articles

A common characteristic of the English and the Hungarian language is the use of "a" article before adjectives and nouns (in Hungarian "a"/"az" pair is used and in English "a"/"an" pair is used). In most case the use of the "a" article does not pose a problem, however, in case of echocardiogram reports, "A" (A wave) is the peak velocity flow in late diastole caused by atrial contraction. Furthermore, in Hungarian language "e" expletive is also present, but in echocardiography reports "E" (E wave) stands for the peak velocity blood flow from gravity in early diastole.

Typographical errors

The lack of a unified recording interface infers many typographical errors which need to be taken into account during term extraction. Most frequent typographical errors can be resolved by using a dictionary which contains the original form of medical terms and their synonyms. If the similarity of the written expression to any term of the dictionary is within an acceptable margin, it is resolvable.

Missing whitespaces

As a form of typographical error, missing whitespaces can also occur between terms, values and units (e.g., EN: "Left ventricular end-diastolic diameter43.: mm", HU: "Bal kamra diast.átm43.: mm"). If the text processing method is word-based, missing whitespaces have a huge impact on the success of processing. This problem can be handled by inserting separator space characters into the text during text

cleaning, if text cleaning is applied.

Recognition of composite terms

Not only typographical errors make it harder to extract information from echocardiography reports. Based on the assumption, that named-entities follow the term-value pair structure, it is possible to extract the greater part of named entities. However, special cases are also present in echocardiography reports mostly because of the habits of the recording individual. Such a composite term can be described in *prefix–term1–term2-value1–value2–common_unit* form (e.g., EN: "left ventricular end-diast/end-syst diameter: 54/35 mm", HU: "bal kamra diast/syst átmérő: 54/35 mm") where the recording individual aggregates two somewhat related terms. In this case the identified term should be interpreted as *prefix–term1–value1–unit* and *prefix–term2–value2–unit*.

Other composite terms can also be present. For example EN: "ejection fraction Teichholz: $56\,\%$, Simpson $52\,\%$", HU: "ejekciós frakció Teichholz: $56\,\%$, Simpson: $52\,\%$" can be described as *prefix–term1–value1–unit–term2–value2–unit* or "E/A: 0.4/0.8 m/s" can be describe as *term1–term2–value1–value2–unit*.

Furthermore, expletives are also commonly used (e.g., EN: "left atrium: 42 mm (apical 4Ch: 43x75mm)", HU: "Bal pitvar: 42 mm (csúcsi nézetből: 43x57 mm)") which makes composite term recognition even harder. A possible approach to process composite terms is to define some basic rules and process echocardiography documents based on these rules.

3.2 Evaluation of different text similarity metrics

To develop a text-similarity-based information extraction method, exact matching is not a viable option, as it is not capable of finding synonyms, typos, and abbreviations of the search term. For this purpose, I examined and compared different text similarity metrics applied in the field of NLP. My goal was to determine which similarity metrics present the highest gain in terms of searching for echocardigoraphy documents containing a given keyword or its misspelled, abbreviated or synonym form.

3.2.1 Included metrics

The basic distance metrics included in my case study are widely used metrics in the field of NLP. The metrics of the study were the following: Longest Common Subsequence (LCS), Levenshtein distance (LD), weighted Levenshtein distance (WLD), Jaro-Winkler distance, and cosine distance. In the following, these metrics and the principles behind them are introduced in detail.

Longest Common Subsequence

Longest Common Subsequence (LCS) is one of the simple metrics measuring the similarity of two strings. It finds the longest subsequence of characters present in both texts. To measure the similarity of the two strings, the actual common subsequence is irrelevant, only the length of it is taken into account [125]. For example, both "cardi" and "ardil" are subsequences of "cardiology" and their length in both cases equals to five. The term subsequence is defined as follows. Given a sequence $a = a_1, \ldots, a_k$. Another $b = b_1, \ldots, b_m$ sequence is a subsequence of a if such a strictly increasing sequence of indices $(i_1, \ldots, i_m)$ of a exists that for all $j = 1, \ldots, m$, $a_{i_j} = b_j$. This metric also takes the cases into account where some characters are omitted, but it cannot recognise swapped characters.

Levenshtein distance

The Levenshtein distance [126, 127] is a more complex dissimilarity metric that counts the number of the edits that are needed to transform an s_1 string into another s_2 string. The Levenshtein distance takes the following operations into account: insertion, deletion, and substitution of characters. The Levenshtein distance works basically on single words, however, it is not restricted to those: it can also be calculated for strings of any type.

$lev_{s_1,s_2}(|s_1|, |s_2|)$ denotes the Levenshtein distance of strings s_1 and s_2, where $|s_1|$ and $|s_2|$ yield the lengths of strings s_1 and s_2, and $lev_{s_1,s_2}(i, j)$ for each $i, j \in \mathbb{N}$ is calculated as

$$lev_{s_1,s_2} = \begin{cases} max(i,j) & \text{if } min(i,j) = 0 \\ min \begin{cases} lev_{s_1,s_2}(i-1,j) + 1 \\ lev_{s_1,s_2}(i,j-1) + 1 \\ lev_{s_1,s_2}(i-1,j-1) + 1_{(s_{1i} \neq s_{2j})} \end{cases} & \text{otherwise} \end{cases}. \quad (3.1)$$

The inclusion of the Levenshtein distance was motivated by the fact that in medical texts the Latin medical terms are probably written, to some degree, in a way similar to spoken-language, and this kind of difference between two words can be easily caught by the use of the Levenshtein distance.

Weighted Levenshtein distance

The original Levenshtein distance is not flexible enough to consider the magnitude of errors: all edit operations uniformly cost 1. However, practically, not all edits can be considered equivalent. For example, in case of typo correction substituting "r" for "t" should have a smaller cost, since they are located close to each other on a keyboard with QWERTY layout. The weighted Levenshtein distance considers all these aspects as well and sets different costs to the pairs of characters according to the probability of their interchange.

Jaro-Winkler distance

The Jaro-Winkler distance [127] accounts for the lengths of two strings and partially accounts for the type of typographical errors humans make when typing texts. The Jaro-Winkler distance is calculated as

$$d_w(s_1, s_2) = 1 - sim_w(s_1, s_2), \tag{3.2}$$

where

$$sim_w(s_1, s_2) = sim_j(s_1, s_2) + lp(1 - sim_j(s_1, s_2)) \tag{3.3}$$

and sim_j is the Jaro similarity for s_1 and s_2 strings, l is the length of a maximum 4 characters long common prefix and p is a constant scaling factor with a standard value of 0.1. The Jaro similarity (sim_j) is calculated as

$$sim_j(s_1, s_2) = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}, \tag{3.4}$$

where $|s_i|$ is the length of s_i, m is the number of matching characters and t is half of the number of transpositions. The concept of matching and transpositions is detailed in [127].

The Jaro-Winkler distance metric results in smaller distance values for those two strings that match from the beginning in length l. I decided to analyse the

applicability of this kind of distance metric as well, because my hypothesis based on the manual review of a pre-selected sample document was that typing errors are more common toward the end of the words.

Cosine similarity

The cosine similarity is also a widely used similarity metric for comparing two strings. For example, it was used in anomaly detection in web documents [128], in content-based recommender systems [129], and even it was used for pattern recognition in medical diagnoses [130]. In the case of calculating cosine similarity, the strings are represented as vectors, and the similarity is calculated from the angle enclosed by the vectors. More formally, the cosine similarity is defined as the inner product of two vectors divided by the product of their lengths. To get the cosine similarity of two strings, the compared strings first have to be projected to a high-dimensional (typically several hundred dimensions) vector space. I achieved it by applying word embedding.

Word embedding [131] is one of the most popular representations of document vocabulary as it is capable of capturing the context of a word in a document, semantic and syntactic similarities or even relations between words. It provides an efficient representation in which similar words have similar encodings. As a result, the words that occur in a similar context will be represented as similar high-dimensional vectors and they tend to have high cosine similarity, as well.

I used the FastText word embedding library developed by Facebook AI Research (FAIR) team to calculate the high-dimensional vector representations for words occurring in medical texts. FastText is an extension of the Word2Vec model proposed by Google [131]. It uses a two-layer neural network for high-dimensional representation. The input of FastText is the word to be mapped with the surrounding text and the output is a high-dimensional representation of the word. The key difference between Word2Vec and FastText is the use of n-grams: Word2Vec only learns from complete words found in the training corpus, while FastText not only considers the complete words, but also the n-grams that are found within each word. The used FastText model was fine tuned manually for the dataset based on the findings of Balázs Szolár [132].

Having the high-dimensional representations of the strings to be compared, the

cosine similarity can be calculated.

$$sim(\mathbf{S}_1, \mathbf{S}_2) = \frac{\sum_{i=1}^{n} s_{1i} s_{2i}}{\sqrt{\sum_{i=1}^{n} s_{1i}^2} \sqrt{\sum_{i=1}^{n} s_{2i}^2}}, \tag{3.5}$$

where vectors $\mathbf{S}_1 = [s_{11}, s_{12}, \ldots, s_{1n}]$ and $\mathbf{S}_2 = [s_{21}, s_{22}, \ldots, s_{2n}]$ are the high-dimensional vector representations of strings (in my case, words) s_1 and s_2.

The examination of cosine similarity was based on the fact that although the same condition may be formulated differently in medical descriptions, but the text surrounding the condition is most likely similar. My hypothesis was that the different descriptions of the same medical terms (e.g., Latin and Hungarian forms of the same term) will get high cosine similarity value. As the similarity of the synonyms cannot be expressed by the application of the similarity metrics presented before, I had great hopes for cosine similarity.

3.2.2 Evaluation process

I had to select the most appropriate text similarity metric as my main goal was to develop a text-similarity-based information extraction method. This choice was made by processing the results of the evaluation process presented in Figure 3.2. The evaluation was done on the corpus presented in Section 3.1. The number of unique expressions found in the original 20,074 reports was 25,380.

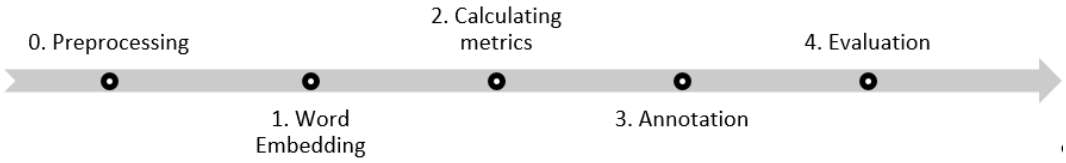

Fig 3.2. Workflow of the evaluation.

All echocardiography reports have been preprocessed as the zeroth step. The aim of *Preprocessing* was to identify the measurements recorded as numerical values and to connect them to their units. The different measured values were replaced with a unified special character to not differ in the preprocessed text. They can be mapped back to their original values at a later step. This way, the variability of the text resulting from different measurement results was significantly reduced, and the number of unique expressions was also reduced to 15,105.

The next step was the design and the execution of *Word Embedding*. The applied FastText-based neural network was tested with different parameter settings. The

best performing neural net utilised a skipgram model with negative sampling. The number of negative samples was 5 and the sampling threshold applied was 10^{-4}. The learning rate was set to 0.05 with a rate of 100 for updating the learning rate. With 20 epochs and a window size of 5, all the words of the preprocessed documents have been mapped to 100-dimensional vectors.

To evaluate the usefulness of the similarity metrics presented in Section 3.2, 10 medical terms considered as important search keywords have been selected by the collaborating doctors' personal preferences. These terms were the following: regurg (insuff), hypokinesia (hypokinézis), akinesia (akinézis), shunt (shunt), bicuspidal (bicuspidalis), thrombus (thrombus), stenosis (szűkület), systolic (systoles), mitral (mitrális), and wallmotion abnormality (falmozgászavar). It is important to note that all these medical terms were given according to the Hungarian terminology, where all of them are expressed with one word. Furthermore, it may seem that the set of selected words is rather small, but in a later step, doctors have to manually annotate the resulting similar words.

In the second step, *Calculating metrics*, the 1,000 nearest matches have been determined for every investigated distance metric and for all of the terms presented in the previous paragraph. All calculated distance values have been converted to similarity values, and the converted values have been normalised to the range of $[0, 1]$. A similarity threshold of 0.65 has also been introduced to limit the number of candidate words.

The third step was *Annotation*. With the help of a cardiologist, a subset of the closest words has been annotated according to the following annotation rules:

- *2*, if the found word was considered identical to the term searched for (e.g., alternative forms, abbreviations, typos)
- *1*, if based on the found word, the cardiologist would consider checking the report containing it to decide whether the report contains relevant information or not, and finally
- *0*, if the found word was considered irrelevant.

The difficulty of the task is shown by the fact that the evaluation of the similar words found for these 10 search words required annotation for 8,647 similar words.

The final step was the *Evaluation*, where the ROC (Receiver Operating Characteristic) curves have been plotted and the corresponding AUC (Area Under the Curve) values have been calculated based on the annotation labels for each search

word: two values for each. The first called *hard evaluation* was calculated where only the words labelled *2* were considered matches, and the second one was the *soft evaluation*, where the words labelled as *1* and *2* were also considered relevant matches.

3.2.3 Results of the evaluation

Table 3.1 shows the resulted number of candidate words in case of setting the similarity threshold equal to 0.65. N_C yields the number of candidates, N_H and N_S the number of the true positive terms for the hard and soft evaluations respectively. The corresponding AUC values for the results presented in Table 3.1 can be seen in Table 3.2. The notation "-" means that with this parameter setting, exclusively real positive candidates were selected and therefore the area under the ROC curve could not be calculated. As we can see, the LCS, Levenshtein, and weighted Levenshtein distances are more capable of distinguishing the true positive candidates from the false positive ones. The advantage of using the weighted version of the Levenshtein distance versus the basic one cannot be observed.

Table 3.1. The number of candidate words in case of applying different similarity metrics and evaluation methods, while setting the similarity threshold equal to 0.65.

	Number of candidate words														
	LCS			Levenshtein			weighted Lev.			Jaro-Winkler			Cosine		
Term	N_C	N_H	N_S	N_C	N_H	N_S	N_C	N_H	N_S	N_C	N_H	N_S	N_C	N_H	N_S
regurg	17	15	15	15	15	15	15	15	15	103	25	29	155	54	91
hypokinesia	76	69	73	58	53	55	34	33	34	470	297	316	489	268	276
shunt	6	6	6	6	6	6	11	6	6	167	19	20	466	48	54
bicuspidal	6	2	2	6	2	2	5	2	2	135	3	4	1000	3	3
thrombus	25	25	25	22	22	22	23	23	23	191	41	49	97	42	52
stenosis	7	6	6	6	6	6	6	6	6	237	15	19	1000	15	52
systolic	50	19	39	45	19	34	39	17	27	259	27	83	136	30	89
akinesia	5	4	4	5	4	4	2	1	1	306	35	36	549	40	62
mitral	27	15	15	22	13	13	12	10	10	519	20	25	151	20	23
wallmotion abnormality	33	28	33	23	18	23	19	17	19	269	28	68	220	28	8

Comparing the two similarity metrics that selected more words as candidates, namely the Jaro-Winkler distance and the cosine distance, we can see that the application of the Jaro-Winkler distance is more appropriate in this topic. After manually reviewing the results, I have established, that although the cosine distance can extract completely different word synonyms as well, they appear in the result set with lower similarity values. Consequently, the false positive results appear with greater similarity values in the results set than the completely different forms of synonyms. Based on my findings, I had to reject the former hypothesis that cosine similarity

Table 3.2. AUC values in case of applying different similarity metrics and evaluation methods, while setting the similarity threshold equal to 0.65.

	AUC									
	LCS		Levenshtein		weighted Lev.		Jaro-Winkler		Cosine	
Term	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft
regurg	1.0000	1.0000	-	-	-	-	0.9954	0.9646	0.8894	0.8398
hypokinesia	0.6004	0.7032	0.6038	0.8545	0.1818	-	0.8483	0.8609	0.8657	0.8695
shunt	-	-	-	-	0.9333	0.9333	0.9968	1.0000	0.9041	0.9143
bicuspidal	1.0000	1.0000	0.8750	0.8750	1.0000	1.0000	0.9975	0.9790	0.9997	0.9997
thrombus	-	-	-	-	-	-	0.9865	0.9789	0.9519	0.8850
stenosis	0.6667	0.6667	-	-	-	-	0.9688	0.9138	0.7919	0.7464
systolic	0.9440	0.7389	0.9413	0.8396	0.7701	0.7191	0.9120	0.9371	0.8041	0.5675
akinesia	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8741	0.8507	0.7622	0.6622
mitral	0.9778	0.9778	0.9829	0.9829	0.8000	0.8000	0.9704	0.9670	0.8729	0.8441
wallmotion abnormality	0.9286	-	0.7000	-	0.9118	-	0.8718	0.9606	0.9269	0.9243

based on the applied FastText embedding can significantly improve keyword-based search in medical texts. The results can be explained by the fact that the contexts surrounding synonyms are probably different for those medical texts that use different expressions for the same content.

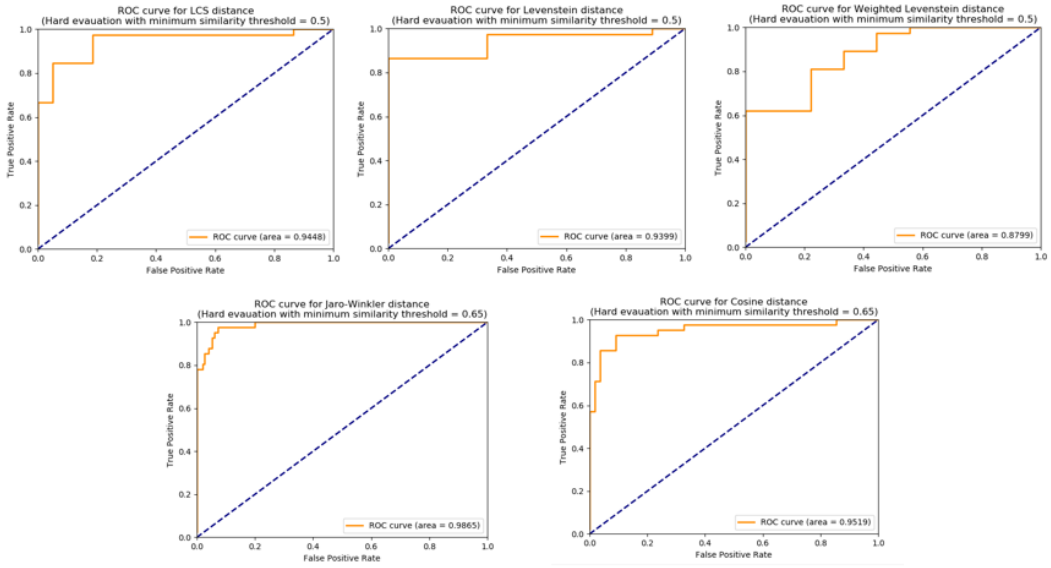

Fig 3.3. ROC analysis of finding the term "thrombus" using different similarity measures.

The results presented in Table 3.1 and Table 3.2 show that for the LCS, Levenshtein and weighted Levenshtein metrics a lower, and for Jaro-Winkler and cosine similarity metrics a higher threshold value has to be applied to get enough true positive candidate words in a way to get good enough AUC value for the classifier. For example, in Figure 3.3, the ROC curves for searching for the word "thrombus" is presented. For finding more positive samples the similarity threshold for LCS, Levenshtein, and weighted Levenshtein distances was decreased to 0.5, while the

threshold for Jaro-Winkler and cosine similarities was left equal to 0.65. The number of the true positive candidates in this case were: $N_{LCS} = 39$, $N_{Levenshtein} = 37$, $N_{wLevenshtein} = 37$, $N_{Jaro-Winkler} = 41$, $N_{Cosine} = 42$. Comparing the results to data listed in Table 3.1 and 3.2, we can see, that as the threshold decreased for the first three similarity metrics, the number of positive candidates increased, however the area under the ROC curve decreased.

My results showed that the Jaro-Winkler and cosine distances, at the same similarity threshold, can discover more candidate words to be similar to the keyword. Although a classifier based on the Common Subsequence, Levenshtein or weighted Levenshtein distances with higher similarity thresholds is more capable of distinguishing the true positive candidates from the false positive ones, with these high thresholds, these metrics provide less true positive results. Furthermore, my results pointed out, that the weighted Levenshtein distance cannot substantially contribute to improving the result of the Levenshtein distance.

Considering the applicability of the cosine distance based on the FastText word embedding, I found that the exploration of synonyms requires a significantly lower threshold, which results in the decrease of the efficiency of the classifier, as well.

The most promising distance metric was the Jaro-Winkler distance, which can return a relatively large number of documents in a way that the distinctive ability of the classifier still remains high. There, the developed text mining-based information extraction method utilises the Jaro-Winkler distance.

3.3 The proposed text mining-based information extraction method

Going beyond the limitations of methods introduced at the beginning of Chapter 3, I developed a generally applicable text mining method for extracting numerical test results with their descriptions from free-text-written echocardiography reports. The proposed method breaks with regex-based information extraction methods and employs corpus-independent text mining techniques to extract information from medical texts. It automatically detects expressions containing textual descriptions of the test results and pairs them with their numerical measurement results. The identification of candidate terms is performed by using similarity-based matching to match them to standardised clinical terms. The similarity-based mapping uses Jaro-Winkler distance and makes it possible to handle typos, synonyms and abbre-

viations flexibly; therefore, the efficiency of the information extraction is significantly increased. Additionally, the proposed method can extract multiple information from the documents by a single search, and a repetitive scan is not needed. The proposed method is mainly recommended for the rapid processing of large volumes of echocardiography findings, such as to support medical research or to verify patient selection criteria for clinical trials quickly.

The applicability of the proposed method was tested by processing the corpus presented on Section 3.1. Figure 3.4 shows how the proposed method extracts and transforms the measurement results of the raw echocardiography document into a uniform and structured form.

Ao. root: 38 mm
Left atrium M-mode 45 mm.
Septum end diastolic 12 mm syst 14 mm Posterior wall end diasolic: 10mm syst 15mm left ventr diast 55 mm, syst 30mm
2D right ventricular diast basal: 40 mm.
ejection fraction (visual estimation) 60 %

Mildly dilated left atrium. Mild concentric left ventricular hypertrophy. Aortic valve is sclerotic, but the opening is normal. Moderate to severe aortic regurgitation. Mild calcified mitral apparatus, mitral valve regurg. gr. I.
Trace tricuspid regurgitation. Pulmonic valve regurgitation: I-II. No obvious wall motion abnormalities.

(a) Echocardiography report

Echo parameter	Value
aortic root	38 mm
M-mode left atrium	45 mm
end-diastolic septum	12 mm
end-systolic septum	14 mm
end-diastolic posterior wall	10 mm
end-systolic posterior wall	15 mm
left vetricle end-diastolic	55 mm
left vetricle end-systolic	30 mm
right ventricle	40 mm
EF	60 %

(b) Extracted result

Fig 3.4. The (a) raw echocardiography report and the (b) extracted measurement results.

3.3.1 Extracting measurement results from echocardiography documents

The steps of the proposed method can be seen in Figure 3.5. The steps are the following: (1) corpus-independent preprocessing of echocardiography documents; (2) identification of the candidate technical terms; (3) refinement of the identified candidate terms; (4) mapping the candidate terms for the standardised clinical terms; (5) validation of the extracted terms and their measured results. The steps are detailed in the next paragraphs.

Preprocessing: The preprocessing phase includes such text cleaning activities

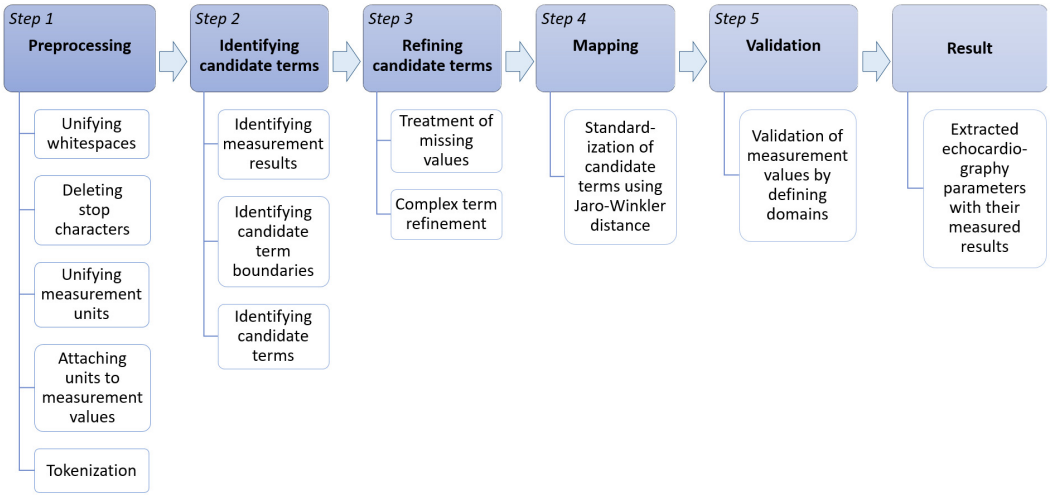

Fig 3.5. The steps of the proposed text mining method.

which aim at unifying the text and minimising the differences between them arising from the recording habits. Therefore, in the first step, the whitespaces are unified and the unneeded characters are deleted according to a predefined character list (*stop characters*). The list of stop characters contains all characters that do not contain any information for the measured results. The list I used and recommend contains the following items: colon, brackets, quotation marks, TAB character and ENTER character. As it can be seen, the dot and comma characters are not the elements of the list because they may also denote decimal separator characters. However, it is important to emphasise that the colon character is the element of the list, and it is deleted at this stage, as the algorithm does not rely on the fact that the measurement results are separated from the name of the measurement by a colon or not.

The next step is to standardise the units of measurement based on a list predefined by an expert. As part of the standardisation, units of measurement containing numbers will be replaced to not contain numbers. (e.g., "mm2" is replaced by "sqrmm"). After standardising the units of the measurements, they are glued to the preceding numerical values (e.g., "85 cm/sec" is replaced by "85cm/sec"). In this way, we can assume that words beginning with a numerical value contain measurement results (typically with their unit of measurement). These "words" are considered *possible measurement results* and are called PMR tokens.

Following this, the cleaned documents are split into tokens, which are text fragments separated by whitespace characters. This tokenised and preprocessed text serves as the basis for the named entity recognition at the next phase.

Identifying candidate terms: This step aims to identify those text fragments that may indicate the names of ultrasound parameters. The proposed method assumes that the description of the measurement precedes the recording of the measurement result. Texts with lengths up to n tokens located between PMR tokens or preceding the first PMR token are considered text fragments that may record the name of a parameter. They are called *echocardiography parameter candidates* (EPCs). The possible maximal number of the words (n) in EPCs is an input parameter of the method and can be established based on the corpus. After the identification, the EPCs and the PMRs are stored in structured form for subsequent processing. We have to note that EPCs do not correspond to the exact names of measurements in all cases; they may still contain complex terms, typos, and abbreviations.

Refining candidate terms: In this phase, complex EPCs will be refined using text fragmentation methods.

If an EPC contains a token describing a unit of measure, it must be divided into two or more parts, since in this case, we can assume that the first part of the term candidate refers to an empty measurement result, while the second part contains another measurement. Given the "A cm/sec EF $62\,\%$" example, the complex EPC contains the "A cm/sec EF" string, which will be cut into two EPCs, which are "A cm/sec" and "EF". The original PMR value ("$62\,\%$") will be connected to the EPC "EF", and the first part of the text will be stored as "A" and "cm/sec" EPC-PMR pair. If the first part contains more than one unit of measurement, then the splitting has to be done recursively in several steps.

The next activity is to recognise and handle the *complex term–measurement sequences*. The forms of the complex sequences may be as follows:

- *term1–term2–result1–result2* sequence: e.g., "left ventricular diameter end-diastolic/end-systolic 54/35mm",
- *term1–result1–subterm2–result2* sequence: e.g., "ejection fraction Teichholz $56\,\%$ Simpson $52\,\%$".

The occurrences of complex sequences are searched using predefined rules. If the sequence fits any of the rules, then the complex sequence is converted into simple term-result pairs, and the refined EPCs with their PMRs are stored.

Mapping: As identified EPCs may contain typos and abbreviations, the next phase of the text processing aims to clarify and standardise them. For achieving

this goal, EPCs are mapped onto a dictionary containing the standardised names of the ultrasound parameters and their synonyms (e.g., "end-diastolic posterior wall", "LVPWd"). The dictionary can be based on any standardised collection of clinical terms (e.g., SNOMED CT [133]), or it can also be defined by experts.

In Section 3.2, I showed that among the different distance metrics, the Jaro-Winkler distance achieves the best results in named entity recognition performed in echocardiography documents. Therefore, in the present method, Jaro-Winkler distance-based mapping is used.

For each EPC, a distance matrix is calculated, which contains the Jaro-Winkler distances between the EPC string and the elements of the standardised dictionary. If the smallest value of the distance matrix is less than a predefined threshold value (α), then the EPC is mapped to the standardised name of the most similar dictionary element. Otherwise, the EPC is discarded.

Validation: During this step, the PMR values are validated by defining a range of interpretations for each ultrasound parameter. If the measured value does not fall within its domain, it will be treated as a possible error.

As a result of the previous steps, the measurement results extracted from the echocardiography documents are available in a structured format (measurement name - measurement result pairs). This structured format allows users (e.g., physicians) to process and use the results in their future work easily (e.g., patient selection for studies or time-series comparison of results).

3.3.2 Evaluation of the proposed text mining-based information extraction method

The effectiveness of the proposed method was evaluated by processing the corpus presented in Section 3.1. During the evaluation, the effectiveness of the extraction of 12 commonly measured echocardiography parameters was examined. These parameters were: aortic root diameter (aorta gyök), M-mode left atrial diameter (Bal pitvar), end-diastolic septum thickness (Septum végdiast), end-systolic septum thickness (Septum syst), left ventricle end-diastolic diameter (Bal kamra diast. átmérő), left ventricle end-systolic diameter (Bal kamra syst. átmérő), end-diastolic posterior wall thickness (Hátsófal végdiast.), end-systolic posterior wall thickness (Hátsófal syst.), right ventricle end-diastolic diameter (Jobb kamra), A-wave (A), E-wave (E), and left ventricular ejection fraction (EF).

The threshold for the maximal number of the tokens in EPCs was set to $n = 4$,

and the threshold parameter for the Jaro-Winkler distance between the EPCs and the standardised terms was set to $\alpha = 0.1$. The first parameter was determined as the suggestion of the medical expert, while the second one was obtained as an empirical fact from my previous study [134]. The cardiac ultrasound documents were processed in a single iteration, and the detected EPCs were mapped into a standardised, expert-defined dictionary. The dictionary used for mapping contained 179 synonyms for 40 terms. The investigated 12 terms had 56 synonyms in total.

After performing the proposed method, the evaluation was carried out the following way. For each document, I examined what measurements the method was able to extract from it. If it was able to extract a given measurement result from a document, the document was tagged with a positive label for that measurement parameter (predicted positive, PP) and if it was unable to extract the given measurement parameter, the document was labelled as negative document for that parameter (predicted negative, PN). The tagging results are shown by Figure 3.6.

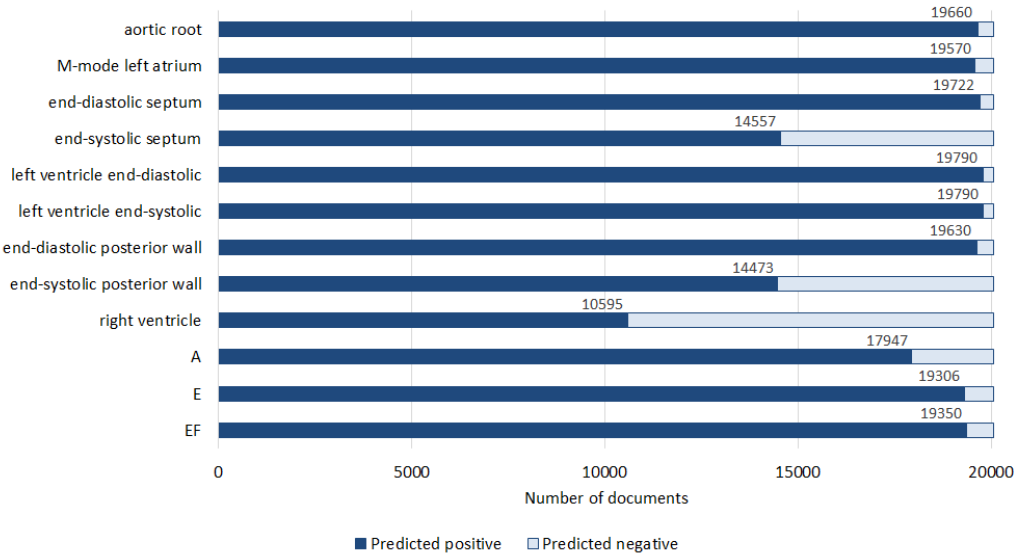

Fig 3.6. Number of documents containing (predicted positive) and not containing (predicted negative) the given term.

Figure 3.6 shows that for eight echocardiography parameters, the algorithm could extract the measurement results from more than 19,000 reports. For these parameters, the relative frequencies were as follows: aortic root diameter: $97.9\,\%$, M-mode left atrial diameter: $97.5\,\%$, end-diastolic septum thickness: $98.2\,\%$, left ventricle end-diastolic diameter: $98.6\,\%$, left ventricle end-systolic diameter: $98.6\,\%$, end-diastolic posterior wall thickness: $97.8\,\%$, E-wave: $96.2\,\%$, and EF: $96.4\,\%$. Relevant information for the remaining four parameters could be extracted from

fewer documents as these parameters are measured less frequently during echocardiography examinations. This was also confirmed by a medical expert. The relative frequencies for these are: end-systolic septum thickness: $72.5\,\%$; end-systolic posterior wall thickness: $72.1\,\%$; right ventricle end-diastolic diameter: $52.8\,\%$; A-wave: $89.4\,\%$.

In the next phase, the quality of the information extraction algorithm was evaluated. 100 reports having predicted positive labels and 100 reports with predicted negative labels have been randomly selected for each investigated term for evaluation purposes. Following this, these 2,400 selected reports have been manually evaluated, and they were labelled with true positive (TP), false positive (FP), true negative (TN) and false negative (FN) labels according to the comparison between the actual content of the document and the prediction. Using these results, numerous evaluation metrics, like sensitivity, specificity, positive predictive value (PPV, precision), negative predictive value (NPV), accuracy, balanced accuracy (Bal. acc.), and F1 score have also been calculated. The values of the calculated metrics can be seen in Table 3.3. All values are rounded to three decimal places.

Table 3.3. Evaluation of the effectiveness of the proposed text mining-based information extraction method.

	#PP	#PN	Sensitivity	Specificity	PPV	NPV	Accuracy	Bal. acc.	F1
aortic root	19660	414	0.877	1.000	1.000	0.860	0.930	0.939	0.935
M-mode left atrium	19553	521	0.813	1.000	1.000	0.770	0.885	0.907	0.897
end-diastolic septum	19722	352	0.862	1.000	1.000	0.840	0.920	0.931	0.926
end-systolic septum	14557	5517	1.000	1.000	1.000	1.000	1.000	1.000	1.000
left ventricle end-diastolic	19790	284	0.926	1.000	1.000	0.920	0.960	0.963	0.962
left ventricle end-systolic	19790	284	0.943	1.000	1.000	0.940	0.970	0.972	0.971
end-diastolic posterior wall	19630	444	0.775	1.000	1.000	0.710	0.855	0.888	0.873
end-systolic posterior wall	14473	5601	1.000	1.000	1.000	1.000	1.000	1.000	1.000
right ventricle	10595	9479	0.980	1.000	1.000	0.980	0.990	0.990	0.990
A	17947	2127	0.855	1.000	1.000	0.830	0.915	0.927	0.922
E	19308	766	0.909	1.000	1.000	0.900	0.950	0.955	0.952
EF	19525	549	0.901	1.000	1.000	0.890	0.945	0.951	0.948
Average			0.904	1.000	1.000	0.887	0.939	0.952	0.948

In Table 3.3, we can see that both specificities and predicted positive values are equal to 1.0 for each parameter. It means that there were no false positive results predicted. The sensitivity of the proposed method takes values from 0.775 to 1.0, and the average sensitivity for the 12 measurements is 0.904. The accuracy of the algorithm varies between 0.855 and 1.0, and the average accuracy is 0.939. The balanced accuracy takes values between 0.888 and 1.0, and its average value is 0.952. The F1 score also shows sufficiently high values, it takes values between 0.873 and 1.0, and its average value is 0.948. The best results were obtained for end-systolic septum thickness and end-systolic posterior wall thickness. Results show that the

most difficult parameter to obtain was the end-diastolic posterior wall thickness. At the same time, the extraction of the measurement result of ejection fraction as an essential diagnostic measure shows good results with 0.901 sensitivity, 1.00 specificity, 0.945 accuracy, and 0.948 F1 score values.

In the next phase of the evaluation, the false negative documents were examined in detail. The errors were classified into two types: (1) the parameter under study is included in the document and its measured value was recorded with numerical values (Err_{num}) or (2) the parameter under study is recorded in the document, but only textual information is given for it (e.g., "end-diastolic posterior wall: can not be measured") (Err_{text}). Table 3.4 shows the occurrence of two different types of error by terms. Findings that contain only textual information on the parameter under study mainly were found for the A-wave. In the case of this type of error (Err_{text}), the algorithm could not be expected to extract the correct information from the document since the proposed method only aims to extract numerical measurement results. However, in a higher proportion of false negative cases, reports also contained numeric measurement results. This is most noticeable for terms end-diastolic posterior wall thickness with 27 occurrences and left atrium diameter with 22 occurrences. Therefore, this error type has been further analysed.

Table 3.4. Frequency of different error types in false negative documents.

	Err_{num}	Err_{text}
aortic root	5	9
M-mode left atrium	22	1
end-diastolic septum	12	3
end-systolic septum	0	0
left ventricle end-diastolic	6	2
left ventricle end-systolic	6	0
end-diastolic posterior wall	27	0
end-systolic posterior wall	0	0
right ventricle	1	1
A	0	17
E	10	0
EF	10	1
Total	99	34

Table 3.5 shows the detailed presentation of the classification of Err_{num} errors according to their causes. The causes of the errors were classified into the following six categories.

- *missing whitespaces*: there were missing whitespaces in the recorded text (e.g., "aortic root: 26mmleft atrium: 47 mm");
- *multiple values*: there were several measurement results recorded together, therefore the term could not be successfully extracted (e.g., "aorta anulus-sinus valsalva-root: 26-36-33 mm");
- *additional text*: the recorded information contained additional text (e.g., "Aortic root: calcified, 26 mm");
- *defined by other parameters*: the measurement result was given in the document as it is equal to another measurement result (e.g., "E=A");
- *wrong order*: the order of the measurement value and its unit was wrong (e.g., "mm7");
- *wrong format*: the recording format of the measured value does not meet the preliminary expectations (e.g., "left atrium: x47 mm", "left atrium 55xmm", "left atrium: 34x33x40 mm").

Table 3.5. Causes of the error type Err_{num} in false negative documents.

	missing whitespaces	multiple values	additional text	defined by other parameter	wrong order	wrong format	total
aortic root	0	0	5	0	0	0	5
M-mode left atrium	0	7	11	0	0	4	22
end-diastolic septum	10	2	0	0	0	0	12
left ventricle end-diastolic	5	0	1	0	0	0	6
left ventricle end-systolic	6	0	0	0	0	0	6
end-diastolic posterior wall	22	2	0	0	3	0	27
right ventricle	0	0	1	0	0	0	1
E	0	0	2	8	0	0	10
EF	0	0	10	0	0	0	10
total	43	11	30	8	3	4	

As it can be seen in Table 3.5, the top three most frequent reasons for classification error were lack of whitespace characters in the text (43 occurrences), measurement values recorded with additional text (30 occurrences), and multiple measurement results recorded together (11 occurrences). However, it is also clear that not only general errors were found but also errors specific to the parameters. For example, the difficulty in extracting the M-mode left atrial diameter is mainly because the used measurement method is not well defined, and in some cases, 2D measurement was applied. Another typical example is the recording of the value of ejection fraction, which is often only given as an estimated value (e.g., "EF:

estimated $68\,\%$"). Knowledge of this and similar facts could greatly facilitate the development of a postprocessing method that could further reduce the number of false negative documents.

3.3.3 Discussion of the results

The main advantage of the suggested method is that it extracts the measurement results of cardiac ultrasound findings by automatically identifying the text fragments describing the measurement names. For this purpose, it uses an expert-defined dictionary and applies text-similarity mappings to identify the unified name of the measurement. In contrast, methods published in the literature typically perform regex-based information extraction. In these methods, the set of regular expressions has to be defined for each measurement parameter separately, which requires IT skills. Furthermore, regular expressions are created as a result of lengthy iterative manual modifications.

To show the effectiveness of the proposed method in terms of finding the measurement descriptions, a comparative analysis between the direct search and the suggested method was performed. All the dictionary elements used in my study were searched in the document set during this analysis, and the evaluation was performed manually on the same test set. As a part of the comparative analysis, I manually checked whether any related numerical measurement results were recorded in the documents. The direct search results are presented in Table 3.6. All values are rounded to three decimal places.

Table 3.6. Evaluation of the effectiveness of direct search.

	#PP	#PN	Sensitivity	Specificity	PPV	NPV	Accuracy	Bal. acc.	F1
aortic root	19708	366	0.902	0.843	0.830	0.910	0.870	0.872	0.865
M-mode left atrium	19853	221	0.941	0.826	0.800	0.950	0.875	0.884	0.865
end-diastolic septum	14870	5204	0.521	1.000	1.000	0.080	0.540	0.760	0.685
end-systolic septum	25	20049	0.217	1.000	1.000	0.100	0.280	0.609	0.357
left ventricle end-diastolic	3939	16135	0.511	0.700	0.970	0.070	0.520	0.605	0.669
left ventricle end-systolic	4096	15978	0.146	0.124	0.150	0.120	0.135	0.135	0.148
end-diastolic posterior wall	14858	5216	0.556	0.681	0.850	0.320	0.585	0.618	0.672
end-systolic posterior wall	2	20072	0.020	0.600	0.500	0.030	0.048	0.310	0.038
right ventricle	12070	8004	0.941	0.950	0.950	0.940	0.945	0.945	0.945
A	19978	96	0.023	0.093	0.020	0.104	0.061	0.058	0.021
E	19903	171	1.000	0.806	0.760	1.000	0.880	0.903	0.864
EF	19764	310	0.918	0.809	0.780	0.930	0.855	0.863	0.843
Average			0.558	0.703	0.718	0.463	0.550	0.630	0.581

Comparing the results of Table 3.3 and 3.6, we can see that the proposed method performed better in extracting all measurement descriptions. This is, of course, due to the fact that the proposed methodology also includes a text similarity mapping,

which can improve the results significantly. In the case of the "end-systolic septum" and "end-systolic posterior wall", the direct search has found only a few search results. In these cases, the text similarity mapping and the EPC refinement phases yielded excellent results in the proposed methodology. In the case of the A-wave, the main problem was, of course, caused by the fact that "A" as a search term is part of the dictionary used, at the same time, it is also a definite article in Hungarian. The search for the terms "E" and "EF" also causes similar problems due to the brevity of the measurement terms. However, as these descriptions are indeed included as measurement descriptions in several documents, the discrepancy is less significant. Moreover, we have to note that the aim of the direct search was only to find the elements of the dictionary in the echocardiography documents, but the related measurement results were extracted manually. In contrast, the proposed method can find the measurement descriptions and extract the related measurement results. By manual refinement, the regex search could be refined, but as mentioned before, this requires the overview of a huge part of the documents, which is a time-consuming task and results in only corpus-dependent regex terms.

If we consider the possibility of international comparison, we find that previously published methods often aim to extract the results of only a single measurement parameter. Hence, the complete comparative evaluation of the effectiveness of the proposed methodology is not feasible. While partial comparisons can be made regarding the extraction of a single ultrasound parameter, the evaluation of the results should consider that the aim of each method was typically different. For example, many previous studies aimed to extract the ejection fraction mentions, including numerical and text descriptions. In contrast, my method was designed to extract several parameters; but at the same time, it aimed at only the extraction of the numerical measurement result.

Research presented in [53] aimed at extracting EF values and mentions from echocardiography reports. The analysis was based on a regex search, and to measure the quality of the method, 765 reports were evaluated. For defining regex terms, a set of sample documents were visually analysed to determine the structure of the documents. The initial pattern set of the regex expressions defined directly for extracting EF values from the reports achieved only an F1 score of 0.4387, but after several refinement phases of the regex expressions, the highest F1 score was 0.957. Finally, the sensitivity of the proposed whole system was 0.889, and the positive predicted value was 0.950. In contrast, my method achieved a sensitivity of 0.901, a

positive predicted value of 1.0 and F1 score of 0.948 when extracting EF values. Although the results can not be compared directly due to the issues mentioned before (my method was developed to not directly extract the EF values but simultaneously more echocardiography parameters), it can be seen that my method has achieved good results and is competitive.

In [52], a large number of echocardiography reports (621,856) were analysed, and the aim was to extract both the numerically or text-recorded ejection fraction measurement results. First, the text descriptions were searched based on a set of concepts defined by the experts, and then, the numerical values associated with the keyword were searched first backwards, then forwards, starting from the text description. If a numeric value was not found, predefined text descriptions (e.g., "severe") were also searched and extracted. The quality of the proposed method was evaluated based on 200 randomly selected reports manually. The algorithm (including the textual information extraction as well) got sensitivity=0.950 and PPV$= 0.969$. Nevertheless, it should be noted that in the present case, we are talking about a system based on the definition and application of a large set of regex terms, and it was developed exclusively for the extraction of the ejection fraction and therefore only applicable to it.

In the study presented in [55], the authors aimed to extract not only ejection fraction measurements but also other cardiac function measurements. The developed methodology was based on natural language processing using a dictionary lookup, rules, and patterns. In this study, the developed NLP method was again based on a large set of regex expressions fine-tuned for both the term and value identifications. The proposed method was evaluated not only for echocardiography reports but also for general clinical notes and radiology reports. Their evaluation was based on 100-100 documents of each type of record. The method achieved averaged F1 score of 0.844 and averaged precision of 0.982 regarding the echocardiography datasets, respectively, for the investigated 27 measurements. For my method the average F1 score was 0.948 and the average precision was 1.000. A comparison of the extraction efficiency of the cardiac ultrasound parameters involved in both cases is shown in Table 3.7. Table 3.7 contains only those results which are covered by both studies.

It can be seen that for some parameters, the method published in [55] performed better, while for other parameters, my proposed method gave better results. For the most informative cardiac ultrasound parameter (EF), the sensitivity of my proposed method was exactly 0.1 better than the method proposed in [55], and the PPV values

were identically 1.0 for both methods.

Table 3.7. Comparison of results of my text mining-based information extraction method with the results achieved by the method presented by Patterson [55].

	Results by Patterson [55]		My results	
	sensitivity	PPV	sensitivity	PPV
end-diastolic septum	1.000	0.926	0.862	1.000
left ventricle end-diastolic	0.706	1.000	0.926	1.000
left ventricle end-systolic	1.000	1.000	0.943	1.000
end-diastolic posterior wall	0.842	0.970	0.775	1.000
EF	0.801	1.000	0.901	1.000

Evaluating the results, we can see that the proposed method provides similar or better results than other methods published in the literature. However, those methods typically rely on the use of regex-based expressions. In contrast, my method is simple, does not require the definition of regex expressions and does not rely on any assumptions about the form in which the results are recorded. Moreover, though my proposed method can extract many numeric measurement results, it still achieves similar results when compared to the methods that are designed to extract only one specific measurement result.

Like other methods, my method also has some limitations. First, with the proposed method, only numerical measurement results can be extracted; textual descriptions regarding the measurements cannot. If numerically recorded measurement results are supplemented with textual descriptions (e.g., "Aortic root: calcified, 26 mm"), this additional information remains hidden. Additional text information, unusual recording formats and missing spaces can also cause errors when extracting measurement results from the text. However, it should be noted that these errors do not always cause problems. Furthermore, measurement results recorded by another echocardiography result (e.g., "E=A") will be missing from the extracted results. However, the mentioned limitations can be generally eliminated by extending the proposed general methodology with special rules.

3.3.4 Usage outside the field of healthcare

The proposed method can be used, not only for healthcare-related documents. A precondition of this generalisation is to have a field-specific dictionary that can be used during mapping. Another limitation of the methods is the usage of the Jaro-Winkler distance. Its usage has only been tested on my specific corpus, but it can

be easily swapped out with other metrics or a stacking [135] can also be used.

3.4 Related theses

Thesis 2.1

I examined and compared different text similarity metrics applied in the field of NLP to determine which similarity metrics present the highest gain in terms of extracting medical terms from echocardiography documents. The examined metrics were the following: Longest Common Subsequence, Levenshtein distance, weighted Levenshtein distance, Jaro-Winkler distance and cosine distance. I established that the Jaro-Winkler distance is the most suitable to identify medical terms in echocardiography documents written in Hungarian language.

Thesis 2.2

By utilising the findings of the comparison of different text similarity metrics, I proposed a text mining-based information extraction method to extract numerical measurement results from echocardiography documents. The proposed method performs generally applicable, language-independent text-cleaning preprocessing activities, automatically identifies measurement names and results, and returns them in a structured way. The methodology is also able to identify, correct and unify synonyms, acronyms, and typos. Since the method does not contain any language-dependent implementation elements, it is suitable for processing echocardiography findings written in any language.

The proposed text mining-based information extraction method was evaluated on a document set containing more than 20,000 echocardiography reports. During the evaluation, 12 relevant echocardiography parameters were extracted from the documents. As a result, an average sensitivity of 0.904, an average specificity of 1.0 and an average F1 score of 0.948 were obtained. The evaluation sufficiently demonstrated the broad applicability of the method, also confirmed by the experts.

Related publications

P10 Szabolcs Szekér and Ágnes Vathy-Fogarassy. Application of Text Mining Methods on Unstructured Hungarian Echocardiogram Documents. *Proceed-*

ings of the Pannonian Conference on Advances in Information Technology (PCIT 2019), University of Pannonia, pages 187-193, 2019.

P11 Szabolcs Szekér, György Fogarassy, Károly Machalik, and Ágnes Vathy-Fogarassy. Application of named entity recognition methods to extract information from echocardiography reports. *Studies in Health Technology and Informatics*, Vol. 260, pages 41–48, 2019. (Q3)

P12 Ágnes Vathy-Fogarassy, Szabolcs Szekér, Balázs Szolár, and György Fogarassy. The efficiency of different distance metrics for keyword-based search in medical documents: A short case study. *Studies in Health Technology and Informatics*, Vol. 271, pages 232–239, 2020. (Q3)

P13 Szabolcs Szekér, György Fogarassy, and Ágnes Vathy-Fogarassy. A general text mining method to extract echocardiography measurement results from echocardiography documents. *Artificial Intelligence in Medicine*, 143: 102584, 2023. (D1, IF: 7.5)

Summary

The large amount of information stored in health databases shines a spotlight on the possibilities offered by retrospective clinical studies. However, the processing of large amounts of data sets requires new methods and new algorithms in many cases, since due to the unique nature of the operation of healthcare and the complexity of the human biological system, data mining methods can typically only be applied after area-specific extensions. Advanced data science methods adapted to health care can effectively contribute to the implementation of retrospective clinical studies and can provide a basis for a more thorough understanding of the functioning of the human biological system. This new knowledge can help doctors implement personalised medicine.

The aim of my research was to develop such new healthcare-adapted data science methods and algorithms, which can effectively contribute to the extraction of information from large (sometimes unstructured) healthcare data files and to the discovery of the information hidden in the data.

My research covered the following topics: development of new control group selection methods for retrospective case-control studies; developing new similarity measures for evaluating the results of the control group selection; analysing the effect of missing variables during the control group selection process; and extracting information from large, unstructured healthcare datasets.

I developed such control group selection methods that can be widely used in case-control studies, irrespective of the field of the study. This statement is evidenced by the fact, that in a recent study, Pouwels et al used the WNNEM method to select healthy participants from different sites [136]. The method is also mentioned and applied in the Pachama study [137] and in a dissertation [138] written at the University of Duisburg-Essen. The purpose of the research described in the professional paper was to create a dynamic baseline that algorithmically selects a regulatory area as an appropriate comparative reference for a carbon project, while the dissertation analyses the financial situation of Chile.

I also developed an information extraction method that is able to extract numerical measurement results from the echocardiogram reports, regardless of the language of the document. The method was published only recently, so its application has not been mentioned until now. Howerver, due to the set of tools used, it is suitable for processing echocardiograms in any language, so I am confident that its use can be implemented on a wider scale in the near future. Since the proposed method uses general text mining procedures, its application is not necessarily limited to the processing of echocardiogram reports, but its application in other areas is also conceivable.

Összefoglalás

Az egészségügyi adatbázisokban tárolt nagy mennyiségű információ reflektorfénybe helyezi a retrospektív klinikai vizsgálatok nyújtotta lehetőségeket. A nagy mennyiségű adathalmazok feldolgozása azonban sok esetben új módszereket és új algoritmusokat igényel, mivel az egészségügy működésének egyedi jellege és a humán biológiai rendszer összetettsége miatt a adatbányászati módszerek jellemzően csak területspecifikus kiterjesztések után alkalmazhatók. A továbbfejlesztett és az egészségügyhöz igazított adattudományi módszerek hatékonyan járulhatnak hozzá a retrospektív klinikai vizsgálatok megvalósításához, és alapot adhatnak az emberi biológiai rendszer működésének alaposabb megismeréséhez. Ez az új tudás segítheti az orvosokat az egyénre szabott orvoslás megvalósításában.

Kutatásom célja olyan új, egészségügyhöz adaptált adattudományi módszerek és algoritmusok kidolgozása volt, amelyek hatékonyan hozzájárulhatnak a nagyméretű (esetenként strukturálatlan) egészségügyi adatállományokból történő információkinyeréséhez és az adatok közt rejlő információk feltárásához.

Kutatásom a következő témákat ölelte fel: új kontrollcsoport-kiválasztási módszerek kidolgozása retrospektív eset-kontroll vizsgálatokhoz; új hasonlósági mértékek kidolgozása a kontrollcsoport-kiválasztás eredményeinek kiértékelésére; a hiányzó változók hatásának elemzése a kontrollcsoport kiválasztási folyamat során; és információk kinyerése nagy, strukturálatlan egészségügyi adathalmazokból.

A kidolgozott kontrollcsoport kiválasztási módszerek széles körben alkalmazhatók eset-kontroll vizsgálatokban, a vizsgálati területtől függetlenül. Ezt az állítást bizonyítja, hogy Pouwels és társai legutóbbi tanulmányukban a WNNEM módszert használták, hogy egészséges résztvevőket válasszanak ki különböző telephelyekről [136]. A módszert a Pachama tanulmány [137] és egy, a Duisburg-Essen Egyetemen [138] írt disszertáció is említi és alkalmazza. A szakmai tanulmányban ismertetett kutatás célja egy dinamikus alapvonal létrehozása volt, amely algoritmikusan kiválaszt egy szabályozási területet megfelelő összehasonlító referenciaként egy karbon projekthez, miközben a disszertáció Chile pénzügyi helyzetét elemzi.

A kifejlesztett információkinyerési módszer a dokumentum nyelvétől függetlenül képes numerikus mérési eredményeket kinyerni a szívultrahang leletekből. A módszer publikálása csupán a közelmúltban történt meg, így alkalmazására mindeddig nem érkezett említés. Az alkalmazott eszközkészletből adódóan azonban tetszőleges nyelvű szívultrahang feldolgozására alkalmas, így bízom benne, hogy felhasználása a közeljövőben szélesebb körben is megvalósulhat. Mivel azonban a javasolt módszer általános szövegbányászati eljárásokat használ, így alkalmazása nem feltétlen korlátozódik szívultrahang leletek feldolgozására, hanem egyéb területen történő hasznosítása is elképzelhető.

Bibliography

[1] S Mostafa Rasoolimanesh, Naser Valaei, and Sajad Rezaei. Guideline for application of fuzzy-set qualitative comparative analysis (fsqca) in tourism and hospitality studies. In *Cutting Edge Research Methods in Hospitality and Tourism*, pages 137–156. Emerald Publishing Limited, 2023.

[2] Lin Zeng, Yan-Qing Fu, Yu-Yi Liu, Jin-Shui Huang, Jian-Xin Chen, Jun-Feng Yin, Shan Jin, Wei-Jiang Sun, and Yong-Quan Xu. Comparative analysis of different grades of tieguanyin oolong tea based on metabolomics and sensory evaluation. *LWT*, page 114423, 2023.

[3] Laith Abualigah, Mohamed Abd Elaziz, Ahmad M Khasawneh, Mohammad Alshinwan, Rehab Ali Ibrahim, Mohammed AA Al-Qaness, Seyedali Mirjalili, Putra Sumari, and Amir H Gandomi. Meta-heuristic optimization algorithms for solving real-world mechanical engineering design problems: a comprehensive survey, applications, comparative analysis, and results. *Neural Computing and Applications*, pages 1–30, 2022.

[4] Jae W Song and Kevin C Chung. Observational studies: cohort and case-control studies. *Plastic and reconstructive surgery*, 126(6):2234, 2010.

[5] Sholom Wacholder, Debra T Silverman, Joseph K McLaughlin, and Jack S Mandel. Selection of controls in case-control studies: Iii. design options. *American journal of epidemiology*, 135(9):1042–1050, 1992.

[6] Robert Horton. *Encyclopaedic companion to medical statistics*. John Wiley & Sons, 2011.

[7] Mark A Zschoch. Configurational comparative methods: Qualitative comparative analysis (qca) and related techniques, benoit rihoux and charles ragin, eds., thousand oaks ca: Sage publications, 2009, pp. xxv, 209. *Cana-*

dian Journal of Political Science/Revue canadienne de science politique, 44(3):743–746, 2011.

[8] Sholom Wacholder, Joseph K McLaughlin, Debra T Silverman, and Jack S Mandel. Selection of controls in case-control studies: I. principles. *American journal of epidemiology*, 135(9):1019–1028, Jan 1992.

[9] Sholom Wacholder, Debra T Silverman, Joseph K McLaughlin, and Jack S Mandel. Selection of controls in case-control studies: Ii. types of controls. *American journal of epidemiology*, 135(9):1029–1041, 1992.

[10] Christopher G Pickvance. Four varieties of comparative analysis. *Journal of Housing and the Built Environment*, 16(1):7–28, 2001.

[11] Masao Iwagami and Tomohiro Shinozaki. Introduction to matching in case-control and cohort studies. *Annals of Clinical Epidemiology*, 4(2):33–40, 2022.

[12] Anthony D. Harris, Matthew H. Samore, Marc Lipsitch, Keith S. Kaye, Eli Perencevich, and Yehuda Carmeli. Control-group selection importance in studies of antimicrobial resistance: Examples applied to pseudomonas aeruginosa, enterococci, and escherichia coli. *Clinical Infectious Diseases*, 34(12):1558–1563, 06 2002.

[13] Gaby S Pell, Regula S Briellmann, Chow Huat Patrick Chan, Heath Pardoe, David F Abbott, and Graeme D Jackson. Selection of the control group for vbm analysis: influence of covariates, matching and sample size. *Neuroimage*, 41(4):1324–1335, 2008.

[14] PRP Behar, PJZ Teixeira, JMG Fachel, and Andre C Kalil. The effect of control group selection in the analysis of risk factors for extended spectrum β-lactamase-producing klebsiella pneumoniae infections. a prospective controlled study. *Journal of Hospital Infection*, 68(2):123–129, 2008.

[15] John E Ripollone, Krista F Huybrechts, Kenneth J Rothman, Ryan E Ferguson, and Jessica M Franklin. Implications of the propensity score matching paradox in pharmacoepidemiology. *American journal of epidemiology*, 187(9):1951–1961, 2018.

[16] Paul Moser. Out of control? managing baseline variability in experimental studies with control groups. *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*, 257, 2019.

[17] Thomas D Koepsell and Noel S Weiss. *Epidemiologic methods: studying the occurrence of illness*. Oxford University Press, USA, 2014.

[18] Nicholas P. Jewell. Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72(1):11, 1985.

[19] Sarjinder Singh and Sarjinder Singh. Stratified and post-stratified sampling. *Advanced Sampling Theory with Applications: How Michael 'selected'Amy Volume I*, pages 649–764, 2003.

[20] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[21] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[22] Sang Hyun Shin, Song Cheol Kim, Ki Byung Song, Dae Wook Hwang, Jae Hoon Lee, Dongjoo Lee, Jung Woo Lee, Eunsung Jun, Kwang-Min Park, and Young-Joo Lee. A comparative study of laparoscopic vs open distal pancreatectomy for left-sided ductal adenocarcinoma: a propensity score-matched analysis. *Journal of the American College of Surgeons*, 220(2):177–185, 2015.

[23] Michifumi Tokuda, Seigo Yamashita, Seiichiro Matsuo, Mika Kato, Hidenori Sato, Hirotsuna Oseto, Eri Okajima, Hidetsugu Ikewaki, Masaaki Yokoyama, Ryota Isogai, et al. Clinical significance of early recurrence of atrial fibrillation after cryoballoon vs. radiofrequency ablation–a propensity score matched analysis. *PLoS One*, 14(7):e0219269, 2019.

[24] Ya-Wen Chuang, Shih-Ting Huang, Tung-Min Yu, Chi-Yuan Li, Mu-Chi Chung, Cheng-Li Lin, Chi-Sen Chang, Ming-Ju Wu, and Chia-Hung Kao. Acute pancreatitis risk after kidney transplantation: Propensity score matching analysis of a national cohort. *PloS one*, 14(9):e0222169, 2019.

[25] Felix J Thoemmes and Eun Sook Kim. A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, 46(1):90–118, 2011.

[26] Sophia HJ Hwang and Elise Cappella. Rethinking early elementary grade retention: Examining long-term academic and psychosocial outcomes. *Journal of Research on Educational Effectiveness*, 11(4):559–587, 2018.

[27] Di Xu, Sabrina Solanki, and Ashley Harlow. Examining the relationship between 2-year college entry and baccalaureate aspirants' academic and labor market outcomes: Impacts, heterogeneity, and mechanisms. *Research in Higher Education*, 61:297–329, 2020.

[28] Jonathan E Shipman, Quinn T Swanquist, and Robert L Whited. Propensity score matching in accounting research. *The Accounting Review*, 92(1):213–244, 2017.

[29] Michael J Peel and Gerry Makepeace. Propensity score matching in accounting research and rosenbaum bounds analysis for confounding variables. *Available at SSRN 1485734*, 2009.

[30] David O Cushman and Glauco De Vita. Exchange rate regimes and fdi in developing countries: A propensity score matching approach. *Journal of International Money and Finance*, 77:143–163, 2017.

[31] Michael Rosholm, Mai Bjørnskov Mikkelsen, and Michael Svarer. Bridging the gap from welfare to education: Propensity score matching evaluation of a bridging intervention. *Plos One*, 14(5):e0216200, 2019.

[32] Onur Baser. Too much ado about propensity score models? comparing methods of propensity score matching. *Value in Health*, 9(6):377–385, 2006.

[33] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[34] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[35] Giuseppe Biondi-Zoccai, Enrico Romagnoli, Pierfrancesco Agostoni, Davide Capodanno, Davide Castagno, Fabrizio D'Ascenzo, Giuseppe Sangiorgi, and Maria Grazia Modena. Are propensity scores really superior to standard multivariable analysis? *Contemporary clinical trials*, 32(5):731–740, 2011.

[36] J Pearl. Remarks on the method of propensity score. *Statistics in medicine*, 28(9):1415–6, 2009.

[37] Mohammad Ali Mansournia, Nicholas Patrick Jewell, and Sander Greenland. Case–control matching: effects, misconceptions, and recommendations. *European journal of epidemiology*, 33:5–14, 2018.

[38] Yizeng He, Soyoung Kim, Mi-Ok Kim, Wael Saber, and Kwang Woo Ahn. Optimal treatment regimes for competing risk data using doubly robust outcome weighted learning with bi-level variable selection. *Computational Statistics & Data Analysis*, 158:107167, 2021.

[39] Fei Wan. Matched or unmatched analyses with propensity-score–matched data? *Statistics in medicine*, 38(2):289–300, 2019.

[40] Peter C Austin, Paul Grootendorst, and Geoffrey M Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, 26(4):734–753, 2007.

[41] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[42] Szabolcs Szekér and Ágnes Vathy-Fogarassy. Weighted nearest neighbours-based control group selection method for observational studies. *Plos One*, 15(7):e0236531, 07 2020.

[43] Szabolcs Szekér and Ágnes Vathy-Fogarassy. Optimized weighted nearest neighbours matching algorithm for control group selection. *Algorithms*, 14(12):356, 2021.

[44] Paul W Mielke and Kenneth J Berry. *Permutation methods: a distance function approach*. Springer, 2007.

[45] Kenneth J Berry, Janis E Johnston, and Paul W Mielke Jr. Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6):527–542, 2011.

[46] Giovanni Fasano and Alberto Franceschini. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 1987.

[47] Szabolcs Szekér and Ágnes Vathy-Fogarassy. Measuring the similarity of two cohorts in the n-dimensional space. In *THE 11TH CONFERENCE OF PHD STUDENTS IN COMPUTER SCIENCE*, page 151, 2018.

[48] Szabolcs Szekér and Agnes Vathy-Fogarassy. How can the similarity of the case and control groups be measured in case-control studies? In *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 000033–000040. IEEE, 2019.

[49] Szabolcs Szekér and Ágnes Vathy-Fogarassy. The effect of latent binary variables on the uncertainty of the prediction of a dichotomous outcome using logistic regression based propensity score matching. In *eHealth*, pages 1–8, 2018.

[50] Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang. Data processing and text mining technologies on electronic medical records: a review. *Journal of healthcare engineering*, 2018, 2018.

[51] Michael Krauthammer and Goran Nenadic. Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6):512–526, 2004.

[52] Fagen Xie, Chengyi Zheng, Albert Yuh-Jer Shen, and Wansu Chen. Extracting and analyzing ejection fraction values from electronic echocardiography reports in a large health maintenance organization. *Health informatics journal*, 23(4):319–328, 2017.

[53] Jennifer H Garvin, Scott L DuVall, Brett R South, Bruce E Bray, Daniel Bolton, Julia Heavirland, Steve Pickard, Paul Heidenreich, Shuying Shen, Charlene Weir, et al. Automated extraction of ejection fraction for quality

measurement using regular expressions in unstructured information management architecture (uima) for heart failure. *Journal of the American Medical Informatics Association*, 19(5):859–866, 03 2012.

[54] Youngjun Kim, Jennifer H Garvin, Mary K Goldstein, Tammy S Hwang, Andrew Redd, Dan Bolton, Paul A Heidenreich, and Stéphane M Meystre. Extraction of left ventricular ejection fraction information from various types of clinical reports. *Journal of biomedical informatics*, 67:42–48, 2017.

[55] Olga V Patterson, Matthew S Freiberg, Melissa Skanderson, Samah J Fodeh, Cynthia A Brandt, and Scott L DuVall. Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC cardiovascular disorders*, 17(1):1–11, 06 2017.

[56] Quinn S Wells, Eric Farber-Eger, and Dana C Crawford. Extraction of echocardiographic data from the electronic medical record is a rapid and efficient method for study of cardiac structure and function. *Journal of Clinical Bioinformatics*, 4:1–10, 09 2014.

[57] Martin Toepfer, Hamo Corovic, Georg Fette, Peter Klügl, Stefan Störk, and Frank Puppe. Fine-grained information extraction from german transthoracic echocardiography reports. *BMC medical informatics and decision making*, 15:1–16, 2015.

[58] Siddhartha R Jonnalagadda, Abhishek K Adupa, Ravi P Garg, Jessica Corona-Cox, and Sanjiv J Shah. Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of hfpef patients for clinical trials. *Journal of cardiovascular translational research*, 10(3):313–321, 06 2017.

[59] Vinaitheerthan Renganathan. Text mining in biomedical domain with emphasis on document clustering. *Healthcare informatics research*, 23(3):141–146, 2017.

[60] Szabolcs Szekér, György Fogarassy, and Ágnes Vathy-Fogarassy. A general text mining method to extract echocardiography measurement results from echocardiography documents. *Artificial Intelligence in Medicine*, 143:102584, 2023.

[61] Soko Setoguchi, Sebastian Schneeweiss, M Alan Brookhart, Robert J Glynn, and E Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555, 2008.

[62] Peng Zhao, Xiaogang Su, Tingting Ge, and Juanjuan Fan. Propensity score and proximity matching using random forest. *Contemporary clinical trials*, 47:85–92, 2016.

[63] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[64] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

[65] S Cavuto, F Bravi, MC Grassi, and Giovanni Apolone. Propensity score for the analysis of observational data: an introduction and an illustrative example. *Drug Development Research*, 67(3):208–216, 2006.

[66] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

[67] Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

[68] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

[69] Wang-Sheng Lee. Propensity score matching and variations on the balancing test. *Empirical economics*, 44:47–80, 2013.

[70] Peter C Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069, 2014.

[71] Paul R Rosenbaum and Paul R Rosenbaum. *Overt bias in observational studies*. Springer, 2002.

[72] Paul R Rosenbaum, P Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

[73] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political analysis*, 27(4):435–454, 2019.

[74] Peter C Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161, 2011.

[75] Yongji Wang, Hongwei Cai, Chanjuan Li, Zhiwei Jiang, Ling Wang, Jiugang Song, and Jielai Xia. Optimal caliper width for propensity score matching of three treatment groups: a monte carlo study. *PloS one*, 8(12):e81045, Nov 2013.

[76] Myoung-jae Lee. *Matching, regression discontinuity, difference in differences, and beyond*. Oxford University Press, 2016.

[77] Peter C Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, 27(12):2037–2049, 2008.

[78] Markus C Elze, John Gregson, Usman Baber, Elizabeth Williamson, Samantha Sartori, Roxana Mehran, Melissa Nichols, Gregg W Stone, and Stuart J Pocock. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, 69(3):345–357, 2017.

[79] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[80] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[81] Thomas W MacFarland, Jan M Yates, Thomas W MacFarland, and Jan M Yates. Mann–whitney u test. *Introduction to nonparametric statistics for the biological sciences using R*, pages 103–132, 2016.

[82] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[83] Peter C Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009.

[84] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[85] Andrej N Kolmogorov. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.

[86] Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

[87] Harold Hotelling. The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378, 08 1931.

[88] Junyong Park and Deepak Nag Ayyala. A test for the mean vector in large dimension and small samples. *Journal of Statistical Planning and Inference*, 143(5):929–943, 2013.

[89] Muni S Srivastava, Shota Katayama, and Yutaka Kano. A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358, 2013.

[90] Lan Wang, Bo Peng, and Runze Li. A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669, 2015.

[91] Jake Bowers, Mark Fredrickson, and Ben Hansen. Ritools: Randomization inference tools. *R package version 0.1-11*, 2010.

[92] Paul R Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.

[93] William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

[94] Donald B Rubin. Multivariate matching methods that are equal percent bias reducing, i: Some examples. *Biometrics*, pages 109–120, 1976.

[95] Donald B Rubin. Bias reduction using mahalanobis-metric matching. *Biometrics*, pages 293–298, 1980.

[96] Peter C Austin. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in medicine*, 30(11):1292–1301, 2011.

[97] Peter JM Van Laarhoven, Emile HL Aarts, Peter JM van Laarhoven, and Emile HL Aarts. *Simulated annealing*. Springer, 1987.

[98] Raghunath Arnab. *Survey sampling theory and applications*. Academic Press, 2017.

[99] David Hankin, Michael S Mohr, and Kenneth B Newman. *Sampling theory: For the ecological and natural resource sciences*. Oxford University Press, USA, 2019.

[100] Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Antonio Luna, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. Covid-19 detection in radiological text reports integrating entity recognition. *Computers in Biology and Medicine*, 127:104066, 2020.

[101] Chaochen Wu, Guan Luo, Chao Guo, Yin Ren, Anni Zheng, and Cheng Yang. An attention-based multi-task model for named entity recognition and intent analysis of chinese online medical questions. *Journal of Biomedical Informatics*, 108:103511, 2020.

[102] Hanyin Wang, Yikuan Li, Seema A Khan, and Yuan Luo. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artificial intelligence in medicine*, 110:101977, 2020.

[103] Luke T Slater, William Bradlow, Dino FA Motti, Robert Hoehndorf, Simon Ball, and Georgios V Gkoutos. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. *Computers in biology and medicine*, 130:104216, 2021.

[104] Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. Multi-domain clinical natural language processing with

medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083, 2021.

[105] Bethany Percha. Modern clinical text mining: a guide and review. *Annual review of biomedical data science*, 4:165–187, 2021.

[106] Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, pages 466–471, 1996.

[107] Sarah Cohen, Anne-Sophie Jannot, Laurence Iserin, Damien Bonnet, Anita Burgun, and Jean-Baptiste Escudié. Accuracy of claim data in the identification and classification of adults with congenital heart diseases in electronic medical records. *Archives of cardiovascular diseases*, 112(1):31–43, 2019.

[108] Julia T Fu, Evan Sholle, Spencer Krichevsky, Joseph Scandura, and Thomas R Campion. Extracting and classifying diagnosis dates from clinical notes: a case study. *Journal of Biomedical Informatics*, 110:103569, 2020.

[109] Ruchi Sahu. Rule-based method for automatic medical concept extraction from unstructured clinical text. In *Recent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017, Volume 3*, pages 261–267. Springer, Springer, 2018.

[110] João Rafael Almeida and Sérgio Matos. Rule-based extraction of family history information from clinical notes. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 670–675, 2020.

[111] Xiaoyuan Bao, Shuanglian Xie, Kai Zhang, Kai Song, and Yunhaonan Yang. Machine learning based information extraction for diabetic nephropathy in clinical text documents. In *2019 6th International Conference on Systems and Informatics (ICSAI)*, pages 1438–1442. IEEE, 2019.

[112] Irena Spasic, Goran Nenadic, et al. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984, 2020.

[113] Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086, 2021.

[114] Viincenza Carchiolo, Alessandro Longheu, Giuseppa Reitano, and Luca Zagarella. Medical prescription classification: a nlp-based approach. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 605–609. IEEE, 2019.

[115] Natasha Chilman, Xingyi Song, Angus Roberts, Esther Tolani, Robert Stewart, Zoe Chui, Karen Birnie, Lisa Harber-Aschan, Billy Gazard, David Chandran, et al. Text mining occupations from the mental health electronic health record: a natural language processing approach using records from the clinical record interactive search (cris) platform in south london, uk. *BMJ open*, 11(3):e042274, 11 2021.

[116] Natalia Viani, Riley Botelle, Jack Kerwin, Lucia Yin, Rashmi Patel, Robert Stewart, and Sumithra Velupillai. A natural language processing approach for identifying temporal disease onset information from mental healthcare text. *Scientific Reports*, 11(1):757, 01 2021.

[117] Jianliang Yang, Yuenan Liu, Minghui Qian, Chenghua Guan, and Xiangfei Yuan. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. *Applied Sciences*, 9(18):3658, 2019.

[118] Runjie Zhu, Xinhui Tu, and Jimmy Xiangji Huang. Utilizing bert for biomedical and clinical text mining. In *Data Analytics in Biomedical Engineering and Healthcare*, pages 73–103. Elsevier, 2021.

[119] Chaofan Li and Kai Ma. Entity recognition of chinese medical text based on multi-head self-attention combined with bilstm-crf. *Mathematical Biosciences and Engineering*, 19(3):2206–2218, 2022.

[120] Ruoyu Zhang, Pengyu Zhao, Weiyu Guo, Rongyao Wang, and Wenpeng Lu. Medical named entity recognition based on dilated convolutional neural network. *Cognitive Robotics*, 2:13–20, 2022.

[121] Peng Chen, Meng Zhang, Xiaosheng Yu, and Songpu Li. Named entity recognition of chinese electronic medical records based on a hybrid neural network and medical mc-bert. *BMC Medical Informatics and Decision Making*, 22(1):1–13, 2022.

[122] Rebecka Weegar, Alicia Pérez, Arantza Casillas, and Maite Oronoz. Recent advances in swedish and spanish medical entity recognition in clinical texts using deep neural approaches. *BMC medical informatics and decision making*, 19(7):1–14, 2019.

[123] Akhil Vaid, Kipp W Johnson, Marcus A Badgeley, Sulaiman S Somani, Mesude Bicak, Isotta Landi, Adam Russak, Shan Zhao, Matthew A Levin, Robert S Freeman, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. *Cardiovascular Imaging*, 15(3):395–410, 2022.

[124] Mathias Kaspar, Caroline Morbach, Georg Fette, Maximilian Ertl, Lea K Seidlmayer, Jonathan Krebs, Georg Dietrich, Leon Liman, Frank Puppe, and Stefan Störk. Information extraction from echocardiography reports for a clinical follow-up study–comparison of extracted variables intended for general use in a data warehouse with those intended specifically for the study. *Methods of Information in Medicine*, 58(04/05):140–150, 2019.

[125] Stefano Lonardi. *String processing and information retrieval*. Springer, 2010.

[126] Frederic P Miller, Agnes F Vandome, and John McBrewster. Levenshtein distance: Information theory. *Computer science, String (computer science), String metric, Damerau*, page 27, 2009.

[127] Jakub Piskorski and Marcin Sydow. String distance metrics for reference matching and search query correction. In *Business Information Systems: 10th International Conference, BIS 2007, Poznan, Poland, April 25-27, 2007. Proceedings 10*, pages 353–365. Springer, 2007.

[128] Menahem Friedman, Mark Last, Yaniv Makover, and Abraham Kandel. Anomaly detection in web documents using crisp and fuzzy-based cosine clustering methodology. *Information sciences*, 177(2):467–475, 2007.

[129] Azene Zenebe and Anthony F Norcio. Representation, similarity measures and aggregation methods using fuzzy sets for content-based recommender systems. *Fuzzy sets and systems*, 160(1):76–94, 2009.

[130] Jun Ye. Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Mathematical and computer modelling*, 53(1-2):91–97, 2011.

[131] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[132] Balázs Szolár. *Szabad szövegek értékelése mélytanuló neurális hálózat segítségével*. PhD thesis, 2021.

[133] Kevin Donnelly et al. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*, 121:279, 2006.

[134] Ágnes Vathy-Fogarassy, Szabolcs Szekér, Balázs Szolár, and György Fogarassy. The efficiency of different distance metrics for keyword-based search in medical documents: A short case study. *Studies in Health Technology and Informatics*, 271:232–239, 2020.

[135] Gautam Kunapuli. *Ensemble Methods for Machine Learning*. Simon and Schuster, 2023.

[136] Petra JW Pouwels, Chris Vriend, Feng Liu, Niels T de Joode, Maria CG Otaduy, Bruno Pastorello, Frances C Robertson, Ganesan Venkatasubramanian, Jonathan Ipser, Seonjoo Lee, et al. Global multi-center and multi-modal magnetic resonance imaging study of obsessive-compulsive disorder: Harmonization and monitoring of protocols in healthy volunteers and phantoms. *International Journal of Methods in Psychiatric Research*, 32(1):e1931, 2023.

[137] Noah Golmant, Martha Morrissey, Carlos Silva, Felix Dorrek, Bernhard Stadlbauer, Rachel Engstrand, and Dick Cameron. Pachama research brief: A description and initial validation of a dynamic baseline for avoided deforestation projects.

[138] Gonzalo Ignacio Durán Sanhueza. *Marginalisation and fragmentation of collective bargaining in Chile. Impacts on workers' power resources and income distribution*. PhD thesis, Dissertation, Duisburg, Essen, Universität Duisburg-Essen, 2022.