KU LEUVEN

ARENBERG DOCTORAL SCHOOL

Faculty of Engineering Science

An Automated Clinical Decision Support System with Better Interpretability

Xi SHI

Supervisor:

Prof. dr. Bart De Moor

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science

October 2021

An Automated Clinical Decision Support System with Better Interpretability

Xi SHI

Examination committee:
Prof. dr. Patrick Wollants, chair
Prof. dr. Bart De Moor, supervisor
Prof. dr. Johan Suykens
Prof. dr. Jesse Davis
Prof. dr. Marjan van den Akker
Prof. dr. Michaela Black
(Ulster University)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science

October 2021

© 2021 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Xi SHI, Kasteelpark Arenberg 10, bus 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

I guess every child has been asked the same question: what do you want to be when you grow up? My answer was "scientist". But when I was 12, I said to myself, if I were Euler, I should have finished the study of algebra by now. Then I had a relatively more "realistic" dream: make money. That's why I did mathematics in my bachelor program and finance as the double degree, worked for 2 years in a consulting company, and accepted the offer of Master in Finance of University of St. Gallen. I was well prepared for an investment banking career. However, just several days before I started the process of Swiss visa application, I suddenly realized that finance is not my real interest. I made a sharp turn from finance to biostatistics, from St. Gallen to KU Leuven, in just one day. I don't know how this happened even myself. The same story happened again two years later when making the decision to do PhD.

It seems that the important decisions in my life were all made on impulse but I never regret. I'm really grateful that destiny pushed me gently to the correct place I belong. I was correct, I am not as talented as Euler, but I am lucky enough to be a researcher and make my own scientific contributions.

Along the way to become a researcher, there are too many people who helped me and provided supports. This thank you list will definitely start with my supervisor, Prof. Bart De Moor, who provided me the opportunity

to become a PhD student and guided me for 4 years. He gave me the freedom to choose my own research direction and supported me firmly even when I had no publications in the third year. Bart also gave me the trust to take the role in management in the EU H2020 project, which is a valuable experience different from many other PhD students. Many thanks for the trust and supports!

I would like to thank my jury members for their help for the last step of the PhD. Prof. Johan Suykens and Prof. Jesse Davis were in my supervisory committee. They assisted in every step and every milestone of the PhD progress and gave me the feedback for the methodology and suggestions for the next step. I also want to thank Prof. Marjan van den Akker. She is the coauthor of my first journal paper and gave me lots of advice and guidance. That was the first achievement of my PhD and really a huge inspiration for me to continue my study. Marjan, I owe you a big, big thank you, for all your help, supports, and guidance! I would like to thank my external jury member, Prof. Michaela Black, the project leader of MIDAS. She was very supportive to give me the freedom to develop the data analytics as I desired, and helped me to apply for necessary resources. Also I would like to thank Prof. Patrick Wollants who acts as the chair of the jury and organized all these procedures.

I would like to thank the colleagues from Academic Center of General Practice, KU Leuven, who worked together with me for Intego. Prof. Gijs Van Pottelbergh and Prof. Bert Vaes are the doctors, professors, and also researchers. I cannot imagine how busy they are, especially during COVID pandemic. But they always timely replied to my emails, provided feedback and suggestions from the perspective of clinicians, and contributed to the papers. Pavlos Mamouris is the other data scientist, he also gave me his own opinions, which were often very helpful. I also want to thank Prof. Rein Vos from Maastricht University, who provided comments and helped me revise the manuscript over and over with prodigious patience. MIDAS

is a really large project, I cannot name everyone here, but I would like to thank everyone for their help and supports. I especially would like to thank Gorka, Debbie, Johnny, and Joseba, not only for the smooth communication and close collaboration, but also for the care and help.

Certainly I want to thank all my dear former and current colleagues in our group! Gorana, my Serbian sister, reliable teammate, and loyal friend. We worked in the same projects, went through difficulties together, and shared each other's success and joy. We had tears and laughters in the past, and thank you so much for giving me so many good memories. You will definitely succeed in the current fight and continue your success to become the next Dr.

Mauricio, thank you for all the encouragements and help! You are always there, ready to provide information and give help. Sometimes I think you are like a friendly NPC in the Novice Area. But if you were an NPC, you would have been the best NPC in any games, as you only give help, provide information, and teach skills, without any requirements or prerequisites!

I also want to thank Tina, she was always one step ahead of me, so she had explored the unknown areas and checked the unclear conditions for us followers. I want to thank Willem, who helped me in MIDAS project when Gorana was on her maternity leave. Thank you Willem, for supporting me using your expertise in software engineering. Marc, thank you for all the advice and suggestions at the start of my research path! And Thibaut, thank you very much for being such a nice office roommate. I also want to thank Wanqiu, for the helpful discussions and conversations on machine learning. Then I want to thank Melanie, Naomi,and Giulia, who worked together with me in different projects. Thank you all for your supports and collaboration. Christof, Bram, Bob, Oliver, Nico, Hannes, Katrien, Lola, thank you for making the friendly and easygoing atmosphere and your laughters during the lunch hours (although I didn't have lunch myself, your laughters made me delightful). And I also want to thank Ida, Aldona, John,

Wim, Elsy and Maarten for their administrative supports.

Last but not least, I want to thank my families and friends. As most of them don't speak English, I would like to thank them in Chinese. 首先我要感谢我的父母。科研的路上必然是布满荆棘的，期间充满了困惑，迷茫，挫败，和失落。爸爸妈妈永远给我最坚定的支持和最充足的信任。尽管他们完全不理解我的课题，但是他们尽最大的可能给我支持和鼓励。我还想感谢所有的亲人们。我人在国外，所有的亲人们都在关心着我的健康，尤其是疫情期间，家人们比我还关注比利时的各种动态。感谢所有亲人们的关心和爱护。最后我想谢谢我所有的朋友们。读博期间我几乎停止了一切的社交活动，几乎消失在了朋友们的社交圈里。尽管如此，还是有很多朋友给我发消息，关心我的近况，给我鼓励，祝福，和帮助。衷心感谢所有人！

Xi Shi

October 2021

Abstract

A clinical decision support system (CDSS) is a system which analyzes health data to provide data-driven solutions or suggestions which can help clinicians or other health professionals make decisions. There are mainly two types of CDSS, i.e. knowledge-based and non-knowledge-based. Knowledge-based CDSS stores a set of pre-defined rules to make decisions following expert medical knowledge, while non-knowledge-based CDSS uses artificial intelligence, machine learning models, or statistical methods, to train the computer programs from past experiences and discover patterns or rules from clinical data. Although a rapid growth in the use of non-knowledge-based CDSS can be observed recently, they are still facing the challenge of interpretability and clinical relevance, as many machine learning methods are black-box methods, making it difficult to give clinically meaningful interpretation for the results. This is the barrier that must be overcome before the widespread of non-knowledge-based CDSS.

This thesis presents a CDSS with better interpretability and clinical relevance, which covers a full process of clinical decision making, including data cleaning, disease risk factors identification, disease diagnosis and prediction, disease sequence analysis, and the system architecture of a CDSS implementation. We proposed an automated data cleaning method which applied the clinical knowledge to clean the electronic health records data collected from daily consultation. This method largely reduced the

manual efforts and improved the time efficiency, without a significant loss in accuracy and completeness. Then we proposed an ensemble feature selection framework to identify risk factors for chronic diseases. The heterogeneous-distributed implementation and the new voting strategy to combine the results led to a set of algorithm-identified risk factors with better interpretability and clinical relevance. We also discussed the prognostic models for chronic disease prediction. The imbalanced data structure and its effects on the predictive performance were not sufficiently discussed in previous studies on prognostic models for disease prediction. We presented a resampling method which could help accurately identify the positive cases and suggested the evaluation measures with a penalty of false positive. Moreover, when studying multimorbidity, we took the sequence of development into consideration for the first time. We proposed Markov chain analysis and Weighted Association Rules Mining, which were never used for the topic of multimorbidity before. Finally, we present the system architecture of a CDSS from MIDAS project, which showed a possibility to integrate data from multiple sources and provide different types of analytic tools to support decision-making.

In summary, we proposed a CDSS that could provide clinically relevant results with better interpretability for clinicians or other health professionals who were not proficient in machine learning methods. The CDSS could assist them in many steps of clinical decision making, such as risk factor selection, disease diagnosis and prediction, and disease sequence exploration. The CDSS was an integrated solution, and automation and privacy preserving were well considered during the development and implementation.

Beknopte samenvatting

Een Clinical Decision Support System (CDSS) is een systeem dat gezondheidsgegevens analyseert om datagestuurde oplossingen of suggesties te bieden die clinici of andere gezondheidswerkers kunnen helpen bij het nemen van beslissingen. Er zijn hoofdzakelijk twee soorten CDSS, namelijk op kennis gebaseerd en niet op kennis. Op kennis gebaseerde CDSS slaat een reeks vooraf gedefinieerde regels op om beslissingen te nemen op basis van medische expertise, terwijl niet-op kennis gebaseerde CDSS kunstmatige intelligentie, machine learning-modellen of statistische methoden gebruikt om de computerprogramma' s te trainen op basis van ervaringen uit het verleden en patronen te ontdekken of regels uit klinische gegevens. Hoewel recentelijk een snelle groei in het gebruik van niet-op kennis gebaseerde CDSS kan worden waargenomen, staan ze nog steeds voor de uitdaging van interpreteerbaarheid en klinische relevantie, aangezien veel machine learning-methoden black-box-methoden zijn, waardoor het moeilijk is om klinisch zinvolle interpretaties te geven. voor de resultaten. Dit is de barriere die moet worden overwonnen voor de wijdverbreide niet-op kennis gebaseerde CDSS.

Dit proefschrift presenteert een CDSS met betere interpreteerbaarheid en klinische relevantie, die een volledig proces van klinische besluitvorming omvat, inclusief het opschonen van gegevens, identificatie van ziekterisicofactoren, ziektediagnose en -voorspelling, analyse van ziektesequentie

en de systeemarchitectuur van een CDSS-implementatie. We hebben een geautomatiseerde methode voor het opschonen van gegevens voorgesteld, waarbij de klinische kennis is toegepast om de gegevens van de elektronische medische dossiers die zijn verzameld tijdens dagelijkse raadplegingen, te reinigen. Deze methode verminderde de handmatige inspanningen aanzienlijk en verbeterde de tijdsefficiëntie, zonder een significant verlies aan nauwkeurigheid en volledigheid. Vervolgens hebben we een ensemble-kenmerkselectiekader voorgesteld om risicofactoren voor chronische ziekten te identificeren. De heterogeen gedistribueerde implementatie en de nieuwe stemstrategie om de resultaten te combineren, hebben geleid tot een reeks door algoritmen geïdentificeerde risicofactoren met een betere interpreteerbaarheid en klinische relevantie. We bespraken ook de prognostische modellen voor het voorspellen van chronische ziekten. De onevenwichtige datastructuur en de effecten ervan op de voorspellende prestatie werden niet voldoende besproken in eerdere studies naar prognostische modellen voor ziektevoorspelling. We presenteerden een nieuwe bemonsteringsmethode die zou kunnen helpen om de positieve gevallen nauwkeurig te identificeren en stelden de evaluatiemaatregelen voor met als straf vals positief. Bovendien hebben we bij het bestuderen van multimorbiditeit voor het eerst rekening gehouden met de volgorde van ontwikkeling. We stelden Markov-ketenanalyse en Weighted Association Rules Mining voor, die nog nooit eerder voor het onderwerp multimorbiditeit werden gebruikt. Ten slotte presenteren we de systeemarchitectuur van een CDSS van MIDAS-project, dat de mogelijkheid toonde om gegevens uit meerdere bronnen te integreren en verschillende soorten analytische tools te bieden om besluitvorming te ondersteunen.

Samenvattend hebben we een CDSS voorgesteld die klinisch relevante resultaten zou kunnen opleveren met een betere interpreteerbaarheid voor clinici of andere gezondheidswerkers die niet bedreven waren in machine learning-methoden. De CDSS zou hen kunnen helpen bij vele stappen van klinische besluitvorming, zoals risicofactorselectie, ziektediagnose en

-voorspelling, en ziektesequentieonderzoek. De CDSS was een geïntegreerde oplossing en bij de ontwikkeling en implementatie werd goed nagedacht over automatisering en privacybescherming.

Abbreviations

ARM	Association rules mining
ATC	The Anatomical Therapeutic Chemical code
AUC	Area under the ROC Curve
BFSMR	Bagging-based feature selection framework integrating MapReduce
CDSS	Clinical decision support system
CI	Confidence interval
CKD	Chronic kidney disease
CKD	Clinical knowledge database (only in Chapter 2)
DoH	Department of Health
EDA	Exploratory data analysis
EHR	Electronic health records
ESRD	End-stage renal disease
FN	False negative
FP	False positive
GP	General practitioner
HIPE	Hospital In-Patient Enquiry
HR	Hazard ratio
ICPC	International Classification of Primary Care
KPI	Key performance indicator
MI	Mutual information
MIDAS	Meaningful Integration of Data, Analytics and Services
NFBC	Northern Finland Birth Cohort data
OR	Odds ratio
PCRS	Primary Care Reimbursement Service
QoD	Quality of Data
TP	True positive
TN	True negative
UI	User interface

Contents

Preface i

Abstract v

Beknopte samenvatting vii

Abbreviations xi

Contents xiii

List of Figures xix

List of Tables xxi

1 Introduction 1

1.1 Context and Background 1

1.1.1 Electronic Health Records 2

1.1.2 Machine Learning for Healthcare 3

1.1.3 Clinical Decision Support System 6

1.2 Clinical Decision Support System Design and Use Cases . . 10

1.2.1 Data Cleaning and Preprocessing 10

1.2.2 Feature Selection 12

1.2.3 Disease Classification and Prediction 13

1.2.4 Sequence Analysis 15

1.2.5 System architecture of CDSS 16
1.3 Research Projects 16
1.3.1 MIDAS Project 16
1.3.2 Intego Project 19
1.4 Thesis Outline 21

2 Automated data cleaning method 25
2.1 Backgrounds 27
2.2 Methods 29
2.2.1 Study Setting 29
2.2.2 Data Descriptions and Issues 30
2.2.3 Clinical Knowledge Database (CKD) 31
2.2.4 Fuzzy Search 32
2.2.5 Outlier detection 35
2.2.6 The whole cleaning framework 37
2.2.7 Evaluation measures 37
2.3 Results 39
2.4 Discussion 40
2.5 Conclusion 46

3 A Feature Selection Framework 47
3.1 Background 49
3.2 Methods 53
3.2.1 Participants and data source 53
3.2.2 Outcome indicator 54
3.2.3 Embedded feature selection methods 56
3.2.4 Experimental setup 59
3.2.5 Evaluation measures 61
3.3 Results 62
3.4 Discussion 64
3.4.1 Model Interpretability and Clinical Relevance 64
3.4.2 Contributions and limitations 69

3.5 Conclusion 72
3.6 Supplementary Materials 72
3.6.1 Data pre-processing 72
3.6.2 Related Work 75

4 Predictive Models for Chronic Disease 79
4.1 Introduction 80
4.2 Materials and Methods 82
4.2.1 Patients 82
4.2.2 ESRD and Risk Factors 83
4.2.3 Data Preprocessing - Resampling Methods 84
4.2.4 Statistical Analysis 85
4.3 Results 87
4.3.1 Study Participants 87
4.3.2 Resampling Methods 87
4.3.3 Estimates of three models 88
4.3.4 Performance evaluation 89
4.3.5 Threshold comparison 91
4.4 Discussion 91
4.4.1 Imbalanced data 91
4.4.2 Evaluation Measures 94
4.4.3 Generalization 95
4.5 Supplementary materials 95
4.5.1 Supplementary methods 95
4.5.2 Supplementary tables 96

5 Development of Multimorbidity over Time 99
5.1 Introduction 100
5.2 Methods 102
5.2.1 Study Design and Participants 102
5.2.2 Selection of Conditions 104
5.2.3 Statistical Analyses 104

5.3 Results . . . 106
5.3.1 Multimorbidity and Prevalence . . . 106
5.3.2 Probabilities of Follow-Up Conditions . . . 107
5.3.3 Weighted Association Rules . . . 109
5.4 Discussion . . . 113
5.4.1 Strengths and Limitations . . . 113
5.4.2 Validity . . . 115
5.5 Supplementary materials . . . 117
5.5.1 Supplementary methods . . . 117
5.5.2 Supplementary tables . . . 122
5.5.3 Supplementary figures . . . 125

6 System Architecture of a Clinical Decision Support System 131
6.1 Introduction . . . 133
6.2 System Architecture . . . 134
6.2.1 Pilots . . . 134
6.2.2 Platform Overview . . . 135
6.2.3 Data Preparation and Harmonization . . . 136
6.2.4 Data Analytics . . . 138
6.2.5 Data visualization . . . 141
6.2.6 Open and Social Data . . . 144
6.3 Results . . . 148
6.4 Discussion . . . 150
6.4.1 Principal Results . . . 150
6.4.2 Comparison with Prior Work . . . 151
6.4.3 Conclusion . . . 151

7 Conclusion 155
7.1 Contributions . . . 155
7.2 Future work . . . 158

Bibliography 163

List of publications **197**

List of Figures

1.1 Examples of Popular CDSSs 7
1.2 The design of the Clinical Decision Support System 11
1.3 MIDAS Framework 17
1.4 A snapshot of the platform presenting the Intego results . . 20
1.5 The thesis outline 22

2.1 A toy example to show how the fuzzy search method works 34
2.2 Example of a data range 36
2.3 The data cleaning framework to build a data warehouse to be used for follow-up studies 38
2.4 The flowchart of Erythrocyte to give an example of the changes in each step 40

3.1 The selection and exclusion criteria for the participants. . . 55
3.2 The main procedures of BFSMR 60
3.3 Visualization of the comparison of selected variables with high scores calculated from 3 voting strategies 66

5.1 Heatmap of the first-order transition probability matrix of a Markov chain including 103 chronic conditions 108
5.2 Visualization of Weighted Association Rules Mining results . 111

5.3 An Example of a Markov Chain Transition Matrix and Visualization 118
5.4 Subgroup analysis – Markov chain analysis - Male 125
5.5 Subgroup analysis – Markov chain analysis - Female 125
5.6 Subgroup analysis – Markov chain analysis - Age group 40-59 years old 126
5.7 Subgroup analysis – Markov chain analysis - Age group 60-74 years old 127
5.8 Subgroup analysis – Markov chain analysis - Age group above 74 years old 128
5.9 Subgroup analysis – WARM - Sex group 129
5.10 Subgroup analysis – WARM - Age group 130

6.1 MIDAS Platform Overview 136
6.2 An overview of the system configured for MIDAS 137
6.3 GYDRA data preparation tool UI 139
6.4 Cross-filter of the Finnish Pilot 141
6.5 MIDAS UI screenshots 143
6.6 147
6.6 Open and Social Analytics 148

List of Tables

2.1 An example of raw data 30
2.2 Example rows in Clinical Knowledge Database 32
2.3 An example of the similarity ratio for the unit check 33
2.4 Completeness and correctness before and after data cleaning (Top 10 variables with the largest volume) 41

3.1 Top 10 features selected from different models based on variable importance 63
3.2 Comparison of predictive performance among different models. 63
3.3 Comparison of the selected variables with high scores calculated from different voting strategies 65
3.4 A full list of features used in the experiment. 78

4.1 Baseline characteristics of CKD patients over 50 years old in Intego (2005-2015) by future progression or no progression to ESRD 87
4.2 Parameter estimates of Logistic Regression and Cox Model to predict the risk of future progression to ESRD 88
4.3 Prediction performance of Logistic Regression and Cox Models 90
4.4 Prediction performance of Cox Models with different thresholds 91
4.5 Full list of risk factors 97

4.6 Comparison of F scores with different parameters to show the effect of the choices of evaluation measures 98

5.1 Prevalence of multimorbidity by gender group and age group 106
5.2 Results of Weighted Association Rules Mining 112
5.3 Distribution of the total patients and multimorbidity patients by year 122
5.4 Selection of 103 chronic diseases with their prevalence in Intego 123
5.5 The important associations from the Markov Chain results . 124

6.1 Research topics for all pilots 135
6.2 Pilot-specific analytics 142
6.3 The background information of the participants of user experience testing 149
6.4 Research topics for all pilots 153

Chapter 1

Introduction

This thesis explores the application of machine learning methods to support clinical decisions, which is in the domain of medical informatics, with a focus on user-centered technology to help solve questions raised by clinicians or health policy makers. The work is presented as a compilation of research publications.

This chapter introduces and contextualizes the research of this thesis. First, the context and background of this work is presented. Second, the algorithms used to develop the clinical decision support system is described. Then the descriptions of the research projects and use cases are listed. Finally, the outline of the structure of this thesis is presented.

1.1 Context and Background

With the rapid development of electronic systems in all areas in most countries, we are getting into a new era of information explosion. Healthcare industry is facing the challenge of the conversion from old-fashioned practice

to the modern, advanced, digital working style. The electronic software, application, system, and platform can be seen almost in every procedure of the healthcare industry, from hospital registry network to electronic prescription, from digital medical images to mobile monitoring applications.

Some new concepts came up as a trend or even gradually develop into a discipline, such as e-health and health informatics. Eysenbach defined e-health as "an emerging field in the intersection of medical informatics, public health and business, referring to health services and information delivered or enhanced through the Internet and related technologies." [57] Health informatics, which applies principles of computer and information science to the advancement of life sciences research, health professions education, public health, and patient care, is argued that it is interchangeable with e-health, as both a broad definition covering electronic/digital processes in health [185].

This thesis is situated in the domain of health informatics, applying machine learning methods in the healthcare context. This section will briefly summarize the context and backgrounds from 3 aspects, i.e. electronic health records (EHR), machine learning in healthcare, and clinical decision support system (CDSS) respectively.

1.1.1 Electronic Health Records

An electronic health record (EHR) is the systematized collection of patient health information in a digital format, which is stored and exchanged electronically and securely through network-connected information systems [70]. These records can be shared across different health care settings, such as primary care, hospitals, emergency health center, specialized healthcare services (e.g. dentistry), and other long-term care facilities. EHRs may include very different types of data, including demographics, diagnosis, medication, laboratory test results and radiological imaging data [82], and

some other information such as medical history, immunization status, vital signs, and billing information [10, 70]. EHRs can be classified based on the International Organization for Standardization definition, which also defines some terminology to describe EHR types, for example, Electronic medical record (EMR), Electronic patient record (EPR), Population health record, etc [70].

The shift towards EHR systems is inevitable, many factors are forcing the acceleration of the widespread use of EHR system [10]. For instance, the increasing healthcare costs, the rising administrative burden, patients' interest of their own personal records, demands of data from all sources to support clinical research, and so on. The advantages of EHRs are absolutely evident. First, EHRs can effectively assist clinical research and provide more evidence-based findings. Second, patients can be better looked after and the quality of care can be improved [105]. The quality has been efficaciously improved regarding not only medical errors and other clinical outcomes, but also user experience and patient participation. Finally, in terms of organizational outcomes, the application of EHR systems can largely reduce healthcare costs, due to better financial management, less unnecessary or duplicated tests, and less manual inputs [105].

All the data used in this thesis is EHR data.

1.1.2 Machine Learning for Healthcare

Due to the rapid development of electronic systems and the abundance of data resources collected from EHRs, the methodology for clinical research is facing an urgent challenge. Traditional statistical models or epidemiology methods show some limitations. For example, it is difficult to apply such methods over large patient population [149]. Therefore, it is becoming common to apply machine learning models to tackle some clinical questions.

Machine learning methods mainly consists of four categories, namely supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [159]. Supervised learning is a learning task which uses the observations with known outcomes or labels to train the model, and then applies the derived model to predict the outcome of new observations. Examples of supervised learning include classification and regression. Semi-supervised learning has both labelled and unlabelled observations in the training set, while none of the observations is labelled for unsupervised learning. Typical unsupervised learning methods are clustering and association rules mining. Reinforcement learning can be viewed as a reward system, which produces multiple outputs and learns to select the best outcome with the maximum reward. The algorithms are trained by rewarding the desirable actions and punishing the undesirable actions.

There are already many applications of machine learning in healthcare, including disease diagnosis, disease prediction, medical image analysis, personalized medicine, etc. One of the most commonly observed applications is the models to predict health risks. This can be a model to label patients with different risk categories [126], or a model to estimate the mortality rate of patients with a certain disease [147]. The aim of the study can be to provide accurate prognostic prediction [26], or to give support for precision medicine [130]. Another use of machine learning is to assist hospital management. For example, methods can be developed to predict the waiting time for the patients, considering factors collected from different sources [159]. Previous studies have demonstrated many advantages of applying machine learning in healthcare, including lower healthcare costs, higher predictive or diagnostic accuracy, better user experience, etc[159].

After applying machine learning technique for all domains for decades, recently interpretable machine learning is gradually gaining attention, which is also of crucial importance in the domain of healthcare [148]. One definition of interpretable machine learning given by Molnar is "the methods

and models that make the behaviour and predictions of machine learning systems understandable to human" [112]. Traditional evaluation measures for machine learning models may be accuracy and AUC, however, it is not trustworthy enough for the users to merely rely on them, especially when the methods are black-box. It has been found that reliable and understandable explanations and interpretations of the machine learning method results are vitally important to convince the healthcare professionals to implement and utilize the machine learning systems [8]. As machine learning has been applied in almost all topics of healthcare, interpretability should be addressed from many different perspectives. For example, the clinicians would like to know the reasons behind the diagnostic model, the patients need the information how an alert is given based on their personal status, and policy makers have to check the logic of a data-driven suggestion for policy decision-making.

There is a sudden growth of applications of interpretable machine learning in healthcare in the past 2 or 3 years. A review in 2020 summarized 19 studies on interpretable predictive models based on global/local or model-specific/model-agnostic interpretablity [169], involving machine learning methods such as decision trees, logistic regression, generalized linear models, generalized additive models, KNN, GNNExplainer, and some techniques to interpret model results. However, the current studies on interpretable machine learning in healthcare are far from sufficient. Therefore, it is expected to explore more machine learning models with better interpretability.

This thesis has a focus of interpretability of the machine learning models in the context of clinical studies and health policy making, thus we would like to give "interpretability" a broader definition, it represents not only the easiness to explain the results, but also the relevance to prior belief and/or the practice in the real world. As it is easier for the professionals to understand if the results of the methods can be shown in a way similar to

the methods or the facts that they are already familiar with, which can be regarded as "the interpretability for the target group".

1.1.3 Clinical Decision Support System

A clinical decision support system (CDSS) is a system which analyzes health data to provide data-driven solutions or suggestions which can help clinicians or other health professionals make decisions. A CDSS enables integrated workflows, aiming to simplify the working process of decision making and giving assistance for healthcare providers. A well-developed CDSS can effectively reduce the manual efforts of health professionals and lower the risk of misdiagnosis or other medical errors.

Knowledge-based VS non-knowledge-based CDSS

There are mainly two types of CDSS, i.e. knowledge-based and non-knowledge-based [175]. Knowledge-based CDSS stores a set of pre-defined rules to make decisions following expert medical knowledge and literature review findings, which normally takes the form of "IF conditions THEN conclusions" [162]. Knowledge-based CDSS consists of three parts, i.e. knowledge base, inference engine, and a mechanism to communicate [158], of which a knowledge base is of crucial value. On the contrary, non-knowledge-based CDSS does not have it, but uses artificial intelligence, machine learning models, or statistical methods, to train the computer programs from past experiences and discover patterns or rules from clinical data [19]. There are some popular examples of CDSS shown in figure 1.1 [158].

Figure 1.1 is not an exhaustive list of all CDSS, it only gave some popular examples by 2015. However, it revealed a fact that knowledge-based CDSS were much more popular than non-knowledge-based CDSS. Despite the

Shahsavarani A.M, et al. Clinical Decision Support Systems (CDSSs)

Table 2. Popular CDSSs

CDSS	manufacturer	Usage Period	Algorithms	layout	Application fields	website
CADUCEUS	Pittsburg University, Pittsburg, PA, USA.	1970's and 1980's	Knowledge-based, Dissociative reasoning	Have separate computer system	Diagnosis of internal diseases, Educational application.	---
Clinical Rules	Digitalis Rx, Amsterdam, Netherlands	Now active	Knowledge-based, Knowledge management, Clinical Rules Engine, G standard MFB, Andere protocollen	Online	Medicine prescription, consumption monitoring.	http://www.clinicalrules.nl/en
DiagnosisPro (Free)	MedTech USA, Inc.،Los Angeles, CA, USA.	Now active	Knowledge-based	Online, computer software, touch phone applet	Diagnosis and differential diagnosis of more than 11 thousand diseases and 30 thousand medical conditions.	http://en.diagnosispro.com/
DxMate (Free)	Infermedica Sp. Warsaw, Poland.	Now active	Knowledge-based	Online	Diagnosis of more than 500 medical condition.	https://dxmate.com/
Dxplain (Free one-month evaluation version)	Massachusets General Hosptial, Boston, MA, USA	Now active	Knowledge-based, pseudo-probabilistic algorithm, Bayesian logic,	Online, computer software	Diagnosis and differential diagnosis of internal diseases, educational application.	http://www.lcs.mgh.harvard.edu/projects/dxplain.html
ESAGIL (Free)	Esagil Institute, New York, NY, USA.	Now active	Knowledge-based, dissociative reasoning,	Online	Diagnosis of diseases according to signs and symptoms, blood and urine test.	http://esagil.org/
ISABEL	Isabel Healthcare Inc. Ann Arbor, MI, USA.	Now active	Knowledge-based, query string, HL-7, XML, API	Online	Diagnosis of all medical diseases, medicine prescription.	http://www.isabelhealthcare.com/home/default
INTERNIST-I/QMR	Pittsburg University, Pittsburg, PA, USA.	1970's and 1980's	Dialog system, pattern recognition, ranking algorithm,	Have separate computer system	Diagnosis of internal diseases, educational application.	---
Litmusdx	Litmusdx Company, Kolkata, India	Active now	Knowledge-based	Online	Diagnosis and differential diagnosis of 11 thousand diseases, presentation of 300 therapeutic protocols, presentation of 50 thousand medicines,200 thousand medicine usage cautions, medical test interpretations, medical files.	www.litmusdx.com
MYCIN	Stanford University, CA, USA.	1970's and 1980's	Knowledge-based, Bayesian networks, graphical models, decision trees	Have separate computer system	Bacteria identification, blood infections identification, medicine prescription, blood clot diseases, educational application.	---
Prescriptor	Digitalis Rx Company, Amsterdam, Netherlands	Now active	Knowledge-based, Knowledge management, matrix models	Online	Medicine prescription, Online access to medical files	http://ww.prescriptor.nl/en www.digitalis.nl
RODIA	Warsaw Medical University, Warsaw, Ploand.	Now active	Non-knowledge-based, Pattern identification, telemedicine, calibration, linear and angular measurement, phantom calibration	Online	Medical imaging, diagnosis, orthopedic problems, monitoring the bone-fracture remedial	w.glinkowski@parser.com.pl
SimulConsult (Free and Purchase versions)	SimulConsult Inc., Chestnut Hill, MA, USA.	Now active	Knowledge-based, Bayesian inference engine, bioinformatics genome annotation, statistical pattern-matching approach	Online	Diagnosis of 5300 diseases especially genetic and neurological.	http://www.simulconsult.com/

Figure 1.1: Examples of Popular CDSSs. Reprinted from Shahsavarani, A. M. et al. "Clinical Decision Support Systems (CDSSs): State of the art Review of Literature." International Journal of Medical Reviews 2 (2015): 299-308.

slow start, a rapid growth in the use of non-knowledge-based CDSS can be observed recently. Based on a systematic literature review, there were 14 studies on clinical decision support systems for chronic diseases by the year of 2020 [163], showing an increasing trend of publication records during 2016-2020. Besides, 60 ML-CDSS were found which were developed to assist infectious disease clinicians [129].

Nevertheless, non-knowledge-based CDSSs are still facing the challenge of interpretability and clinical relevance, as many machine learning methods are black-box methods, making it difficult to give clinically meaningful interpretation for the results. Moreover, without the pre-defined knowledge base to provide reference, the non-knolwedge-based CDSSs try to detect the potential patterns solely based on the data distribution, leading to the risk of bias if the data is of poor quality. Methods which can be self-corrected based on follow-up feedback may help alleviate the potential risks. These are the barriers that must be overcome before the widespread of non-knowledge-based CDSS. This thesis aims to work on methods that can improve interpretability and clinical relevance of CDSS.

CDSS functions

Many functions can be achieved with the use of CDSS, for instance, administrative systems, diagnostic support systems, prescription support systems, and so on [175]. Nowadays, CDSSs are commonly used to guarantee patient safety. They can effectively reduce medication errors by giving alerts for prescriptions which involves drug-drug interactions [72]. Another example of using CDSS to protect patient safety is the reminder systems, such as the automatic monitoring systems in the ICU [55]. Clinicians also use CDSS to assist clinical management. With the automatic check-ups and reminders, the clinical guidelines are better practiced than before [93], and the patients are more timely contacted to follow the care plan [53].

Moreover, CDSS has been widely applied to assist clinical administrative management and clinical documentation.

The CDSSs mentioned above are mostly knowledge-based systems. One of the most important functions that non-knowledge-based CDSS can achieve is to provide diagnostic support. Such CDSS, also called as diagnostic decision support system (DDSS), is defined as "a computer-based algorithm that assists a clinician with one or more component steps of the diagnostic process" [109]. One DDSS may only focus on one specific disease or one step of the diagnostic process. Due to the variety of clinical domains and different steps in the diagnostic process, there are many types of DDSS. The CDSS designed in this thesis can be regarded as one DDSS for chronic diseases.

CDSS effectiveness and adoption challenges

Based on the report from Centers for Disease Control and Prevention (CDC) of the United States, the evidence showed a strong effectiveness of CDSS [121]. This conclusion is also supported by the findings of many systematic reviews [15, 114, 191].

Nevertheless, there are still challenges for adopting CDSS in clinical practice. For instance, DDSS has not been as commonly adopted as other CDSS types, because of reasons such as negative physician perceptions and poor accuracy [175, 18, 20]. Moreover, when looking at a broader range of CDSS, the main reasons leading to low user acceptance of CDSS also included the lack of trust in CDSS and the low clinical relevance [88]. The results from past surveys and literature review suggested that it could improve user acceptance of CDSS if the decision-making process could be transparent and understandable for clinicians. Therefore, this thesis works on some methods with better interpretability and clinical relevance which can be implemented in CDSS.

1.2 Clinical Decision Support System Design and Use Cases

This thesis will discuss about the whole clinical decision process of disease detection and prediction, including selecting risk factors of a disease, making diagnosis and prediction, and predicting future sequence status. There are already many studies using machine learning models to deal with these problems. The aim of the CDSS proposed in this thesis is to improve the interpretability and clinical relevance of the models without affecting the performance and improve the level of automation. Current use cases generally cover four topics, namely data cleaning, feature selection, disease diagnosis and prediction, and sequential prediction. Figure 1.2 presents the design of the whole clinical decision system.

1.2.1 Data Cleaning and Preprocessing

Preprocessing is of crucial importance in most of the clinical project, especially when the datasets are the real-world medical records from the electronic medical system that is used in the daily practice. The original data can have complicated data quality issues or errors caused by manual mistakes or system errors. There are already some discussion about a general scheme or framework to clean the EHR data[49, 108], however, the discussion about automated data cleaning is still not sufficient.

In clinical projects, because of the complexity of the real-world data, the data cleaning process is often demanding with respect to techniques in data cleaning, data processing, and data analytics, and also requires good coding skills. This can be difficult for people without a professional training in data analytics, adding technical barriers for non-technical people, such as clinicians, health policy-makers, and other healthcare related professionals. Moreover, whenever a new project starts, new variables need to be selected

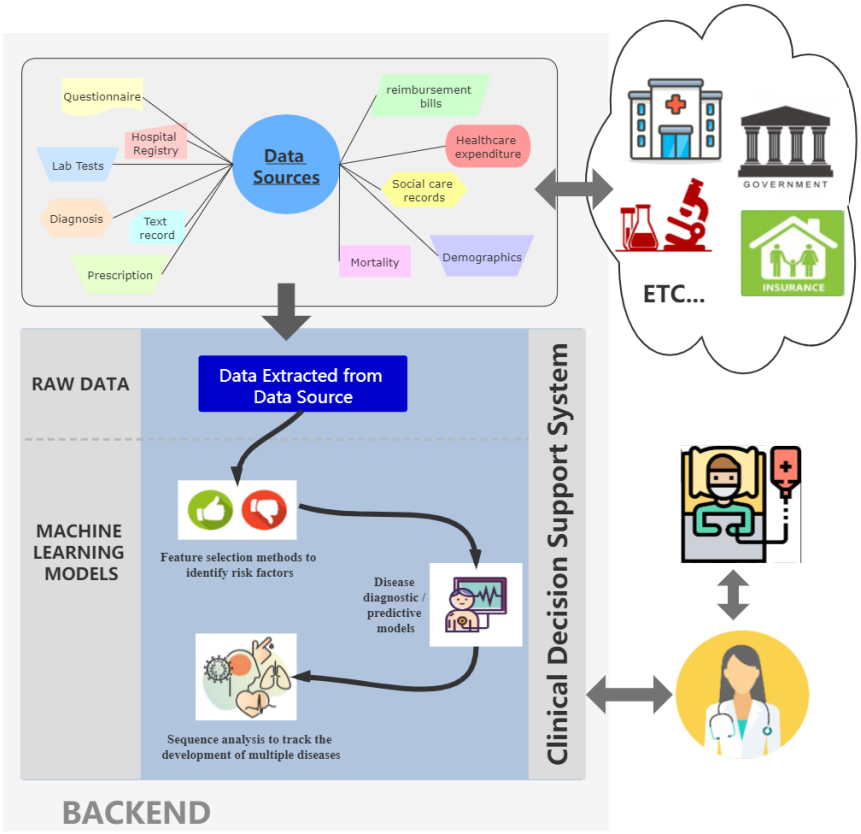


Figure 1.2: The design of the Clinical Decision Support System.

and the data cleaning needs to be done once again. Therefore, an automated data cleaning method can save time and budget spent in this procedure.

Despite the urgent need of the automated data cleaning method, there are not sufficient studies on automated cleaning method for medical data. Most of the papers on data are about Quality of Data (QoD), meaning the dimensions of data quality [194, 59, 178]. With the proposed QoD, some researchers proposed data assessment framework [164, 118] and data cleaning framework [49, 195, 108]. The cleaning strategy is to first identify potential

data issues based on QoD, and then correct the data correspondingly. Extensive manual investigation and decision making is required in many steps. To our knowledge, there was only one paper on automated data cleaning method for EHR data, but it only included one type of variable[133], height and weight.

Case Study: Automated data cleaning method

The EHR data collected in Intego project is the real-world data from daily consultation, meaning that there are typos and mistakes when GPs made data entries or there were data transfer in the system. It takes a considerable amount of time to clean the selected variables whenever a new study is started. It is expected to have an automated data cleaning method, which can finish the data cleaning task without additionally manual input. To achieve the goal, we created a clinical knowledge database to deal with the specific clinical requirements for the data preprocessing and we used fuzzy search to clean the units of the measurements. Finally, we detected the outliers based on clinical knowledge.

1.2.2 Feature Selection

Risk factors need to be selected before disease diagnosis or prediction, as too many predictors may cause over-specification bias. This is the problem of feature selection and interpretability is fatally important in this case. It has to be convincing how the features are selected and why these features are determinants. The results need to be plain and intuitive enough for medical staff and policy makers.

The commonly used feature selection methods select features based on predictive performance and merge the results by using the combination of label prediction, which may lead to a selection of variables with no clinical

relevance and interpretability. The basic rule of this kind of algorithms is that the risk factors are selected if they can lead to the best predictive performance, based on some evaluation measures, such as accuracy and AUC. Hence, the interpretability and clinical relevance of the identified risk factors have been neglected. There are another 2 ways of merging the results, by using feature subsets [101] or using feature ranking [171, 6]. The aim of our work is to propose a feature selection framework with better merging strategy which can select risk factors with good interpretability and clinical relevance.

There is often related issues of big data and high dimensions when working on feature selection. Distributed implementation is one solution when the computation is time-consuming. However, although some studies applied distributed implementation for individual methods [142], few papers tried it for ensemble feature selection [155, 25, 113]. We would like to apply distributed implementation for ensemble feature selection.

Case Study: Ensemble feature selection with good interpretability

We developed a new ensemble feature selection method with distributed implementation by combining the MapReduce with Bagging method. Besides, the way of combining the results are relatively new, which can give a collection of selected features that are easier to be interpreted and implemented in the practice. We used the data of Basque country in MIDAS project and selected the risk factors of childhood obesity. The results proved that the model could give results with better interpretability.

1.2.3 Disease Classification and Prediction

Disease detection and prediction is the problem of classification, and the performance of prediction model largely depends on accuracy. As long as

the model performance is good and the prediction results are accurate, it is not necessary to explain how the model is calculated. In this case, the black-box machine learning models can be applied to give prediction based on patient' s health status using the identified risk factors.

One crucial problem for disease prediction is the imbalanced data structure, as most of the diseases have prevalence lower than 5%, some are even less than 1%. A systematic review on prognostic models for CKD shows that previous studies mostly applied Cox proportional hazards model (87%) or logistic regression (3%) [190], and this is also true for papers on prognostic models for ESRD [143, 95]. However, survival models and logistic regression are unable to efficiently solve the problem of imbalanced data structure. It is necessary to present more models with better performance.

Besides, for clinical classification, there are different preferences for model evaluation measures based on the scenarios. For example, if we are predicting patients' future status for closer monitoring, a wider range of potential patients with healthy people mistakenly involved (False positive) is better than a narrow range without some potential patients. However, if it is to find patients for some practice with side effects, it is better to be more conservative when making diagnosis. Therefore, we need to test more evaluation measures with better clinical relevance.

Case Study: A prognostic model to predict the risk of chronic diseases

End-Stage Renal Disease (ESRD) is the last stage of Chronic Kidney Disease (CKD) and CKD patients with the high risk of ESRD detected at an earlier stage can lower the chance to develop ESRD with good follow-up medical treatment. This requires a good and early identification of high risk patient for an evolution towards ESRD. Previous studies mostly applied Cox model with C statistics or logistic regression with AUC as performance measure. However, we found that the imbalanced data structure caused by

low prevalence and the effects on prediction accuracy were not sufficiently discussed in previous studies. Therefore, we proposed to apply resampling method to preprocess the data before applying the logistic model, which largely improve the ability to detect the true patients. Apart from that, we discussed about the selection of evaluation measures from the clinical perspective.

1.2.4 Sequence Analysis

Based on the literature review, there are six types of models that can be applied for this topic, which are Markov chain, clustering-based models, rule-based models, lossless models, deep learning models, hybrid models, and ARIMA respectively.

The deep learning models have better performance compared to other traditional models, however, it' s difficult to interpret the behavior pattern and how the results are derived. It' s expected to construct a hybrid model which can give good interpretation of the behavior pattern and accurate prediction at the same time.

Case Study: Development of multimorbidity of chronic diseases

The prevalence of multimorbidity is increasing due to aging population and patients with multimorbidity often have a decrease in quality of life and require more health care. The aim of this use case was to explore the evolution of multimorbidity taking the sequence of diseases into consideration, which were rarely done in previous studies. Our study used Markov chains and Weighted Association Rules Mining for the first time to detect frequent combinations of chronic diseases which occur together, with the advantage of taking sequential chronology into consideration.

1.2.5 System architecture of CDSS

All the work discussed above can be combined to develop a CDSS, and we had one use case to discuss one possibility of the system architecture of a CDSS, which presents a solution to how the data can be linked, harmonized, analyzed, and visualized in a multinational framework.

Case Study: System architecture of a healthcare platform

The MIDAS platform delivers a secure, effective and integrated solution to deal with big data to provide support for health policy decision-making, planning of public health activities and the implementation of Health in All Policies. The platform has proved it is transferable, sustainable and scalable across policies, data and regions. The paper on MIDAS system architecture demonstrates how the platform works in the back-end to make data analytics.

1.3 Research Projects

This thesis is presented as a compilation of research publications, including five case studies from two research projects, MIDAS and Intego, covering the full process of a clinical decision support system, from data processing to data analytics, from detailed algorithms to overall platform architecture.

1.3.1 MIDAS Project

The MIDAS (Meaningful Integration of Data, Analytics and Services) project is a European Commission Horizon 2020 project, with a partnership involving health authorities in five EU countries [Basque (Spain), Belgium/Flanders, Finland, Ireland and Northern Ireland (U.K)] and the

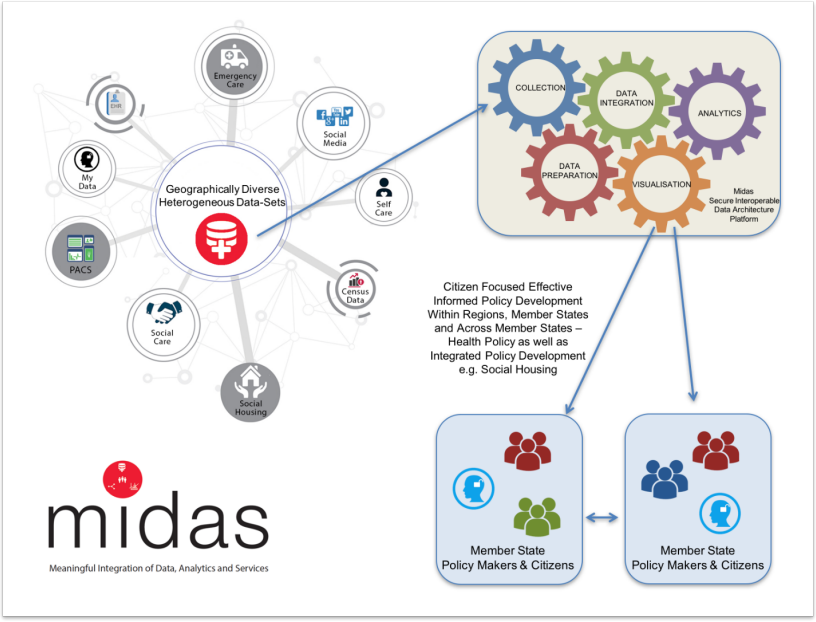


Figure 1.3: MIDAS Framework

U.S. and technical big data experts from research institutions, MNCs and SMEs(Figure 1.3).

The overall aim of MIDAS project is to develop a novel big data platform that connects a range of heterogeneous health-related data with open and social data and applies advanced analytics techniques to provide a visual data-driven decision making tool to enhance healthcare policy making [22]. In more details, the goal of the research project is to develop a platform that is an integration solution enabling:

- Best practice on sourcing, mapping and sharing of existing healthcare data and open data;
- Bespoke visualization of modelling and simulation of the actionable information creating a network of knowledge and;

- Exploit to support public health authorities.

The whole project consists of 8 work packages as below, and we worked in WP4, with a focus on scalable data analytics, which provide the basis to derive insight to support policy makers' decision-making.

- WP1 Project Management and Coordination;
- WP2 Governance, Consent and Privacy;
- WP3 Architecture and Data Sources;
- WP4 Data Analytics and Simulation;
- WP5 Big Data Visualization for Public Health Decision Making;
- WP6 Integration Evaluation and Impact Assessment;
- WP7 Project Dissemination, Exploitation and Awareness Raising;
- WP8 Ethics requirements.

The work presented in this manuscript is mainly related with tools to study the impact of decision making of health policy. To support health policy decision makers, we envision "tools that allows them to simulate the effect of certain policy changes on various aspects of health care, including clinical outcomes, demand, quality of life, and public health economics" [1]. This project also aims to provide supports for simulating the optimal service and making corresponding implementation, by using machine learning methods that were previously tested in previous studies and adapting them to the specific use case.

MIDAS is comprised of four pilots, Northern Ireland, Republic of Ireland, Finland, and the Basque Country (Spain). The Irish dataset comes from Department of Health and a number of divisions of the Health Service Executive. We have data obtained from silo data collection mechanisms,

i.e. HIPE Discharge data and PCRS prescribing data. In the use case of Northern Ireland, Business Services Organization provided data collected from the Health and Social Care Trusts, including information about hospitalization, prescription, and children care. Osakidetza dataset from the Basque Country is composed of data extracted from primary care, specialty care and hospitals databases, including information about patient, diagnosis, medical forms, medical appointments and prescription information and the information about child' s mother. The Finnish datasets include questionnaires data and collected cohort data about mental well-being and physical health information.

The results of the data analytics for Finland, Northern Ireland, and Republic of Ireland are not presented in this manuscript. Two case studies are given in the follow-up chapters, including a feature selection study for Basque (Spain), and the discussion about the system architecture of the MIDAS platform. The feature selection method was developed by the author individually, with the suggestions and feedback from the coauthors to adjust the model, whereas the MIDAS platform (and its system architecture) was developed in collaboration with all WPs in MIDAS project. KUL team worked as data scientists in WP4 Data Analytics and contributed to the development of the analytical engine of the platform. As for the manuscript, the author acted as the leading author to organize the writing and revision.

1.3.2 Intego Project

The Intego project was started in 1990 in Academic Center for General Practice, KU Leuven, with an intention of building a large database on morbidity in primary care to provide information on incidence and prevalence of diseases in Flanders, and on the diagnostic tests and treatments in general practice [120]. Since 1994, the data is yearly collected from the routine data recorded by general practitioners (GP) during the daily

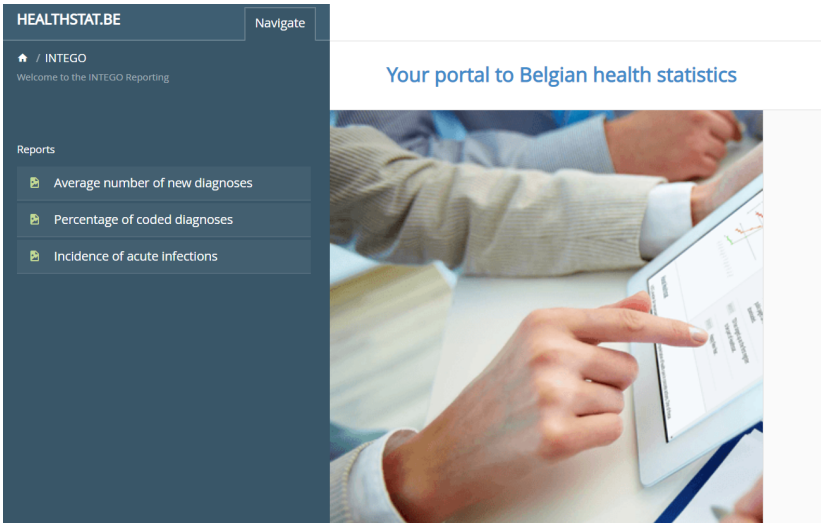


Figure 1.4: A snapshot of the platform presenting the Intego results.

consultation. The Intego database is the first operational computerized morbidity registration network in Belgium based on the electronic health records (EHR) collected from general practice data [184]. Over 90 general practitioners in Flemish areas are involved in the project, and medical information on about 300,000 patients are collected, "corresponding to over 2.3% of the Flemish population representative in terms of age and sex" [184]. Figure 1.4 is a snapshot of the platform which presents part of the analytics results from the Intego project.

There are 6 separate datasets in Intego database, including Patient, Measurements, Lab, Diagnosis, Prescriptions, and Vaccination. Dataset Patient contains patient identity information. Dataset Measurements has variables about measurement codes, measurement dates, measurement names, measurement results, and measurement units apart from patient identity information [160]. This dataset mainly includes the basic health features measured by GPs' commonly used equipment, such as height, weight, heart beat and blood pressure. Dataset Lab has similar structure to

Measurements, and it includes variables collected from blood samples, such as creatinine and cholesterol [160]. Although Diagnosis and Prescriptions have more variables, only diagnosis/prescription codes and dates are relevant [160]. Dataset Vaccination was not used in the thesis. In total, the Intego database contains about 3 million diagnoses, 27 million laboratory results and 12 million prescriptions for medication.

1.4 Thesis Outline

This thesis is presented as a compilation of research publications, organized as indicated in Figure 1.5. Each chapter represents a publication. The publications are only adapted in layout to fit the styling of this thesis. References are merged at the end of this thesis.

The sequence of the following chapters is:

Chapter 2 – Case Study 1: An Automated Data Cleaning Method for Electronic Health Records by Incorporating Clinical Knowledge

This chapter describes a case study of the automated data cleaning method for EHR data. This corresponds to Section 1.2.1 and Section 1.3.2. This chapter has been published by the journal BMC Medical Informatics and Decision Making.

Chapter 3 – Case Study 2: An Ensemble-Based Feature Selection Framework to Select Risk Factors of Childhood Obesity for Policy Decision Making

This chapter describes a case study of an ensemble feature selection method that can provide risk factors of chronic diseases with better interpretability and clinical relevance. This corresponds to Section 1.2.2 and Section 1.3.1. This chapter has been published by the journal BMC Medical Informatics and Decision Making.

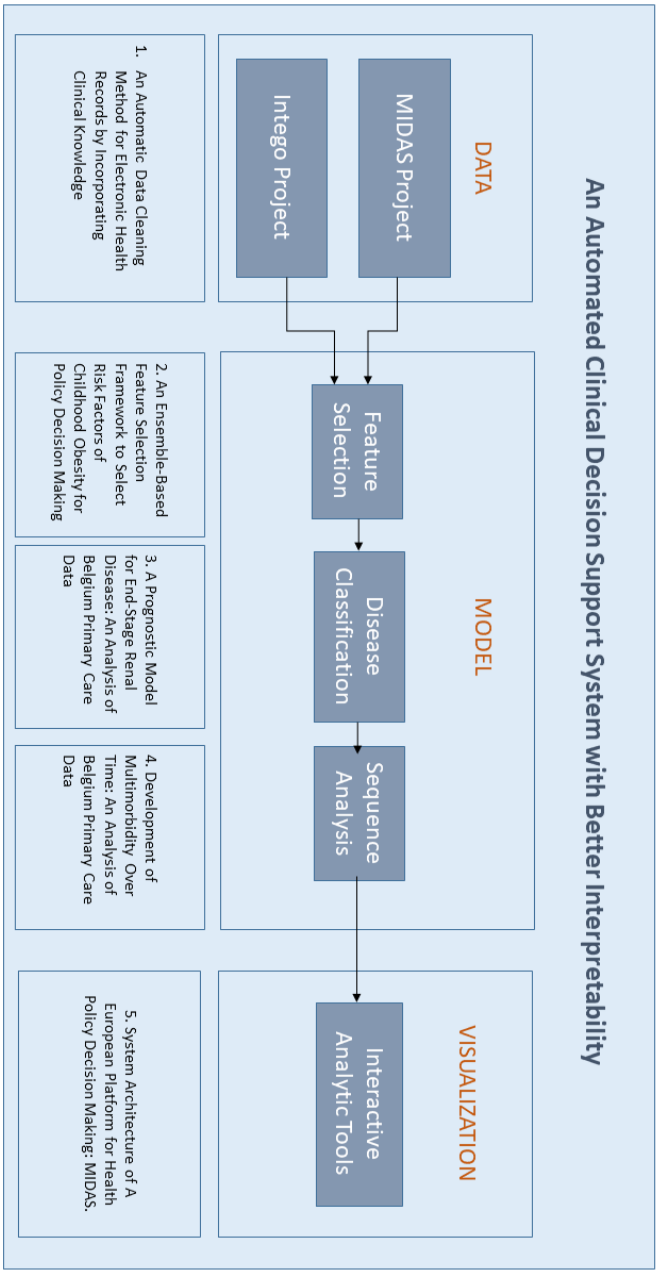


Figure 1.5: The thesis outline.

Chapter 4 – Case Study 3: A Prognostic Model for End-Stage Renal Disease: An Analysis of Belgium Primary Care Data

This chapter describes a case study of a prognostic model that has better performance to predict the positive cases when the data structure is imbalanced. This corresponds to Section 1.2.3 and Section 1.3.2. This chapter has been submitted to the journal Frontiers in Medicine.

Chapter 5 – Case Study 4: Development of Multimorbidity Over Time: An Analysis of Belgium Primary Care Data Using Markov Chains and Weighted Association Rule Mining

This chapter describes a case study of multimorbidity with development sequence taken into consideration. This corresponds to Section 1.2.4 and Section 1.3.2. This chapter has been published by The Journals of Gerontology: Series A.

Chapter 6 – Case Study 5: System Architecture of A European Platform for Health Policy Decision Making: MIDAS

This chapter describes a case study of the system architecture of the MIDAS platform. This corresponds to Section 1.2.5 and Section 1.3.1. This chapter has been submitted to the Journal of Medical Internet Research.

Chapter 2

Automated data cleaning method

This chapter has been published by BMC Medical Informatics and Decision Making as:

Shi X, Prins C, Van Pottelbergh G, Mamouris P, Vaes B, De Moor B. (2021). An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge. BMC Medical Informatics and Decision Making:21:267.https://doi.org/10.1186/s12911-021-01630-7

Abstract

Background: The use of Electronic Health Records (EHR) data in clinical research is incredibly increasing, but the abundancy of data resources raises the challenge of data cleaning. It can save time if the data cleaning can be done automatically. In addition, the automated data cleaning tools for data in other domains often process all variables uniformly, meaning

that they cannot serve well for clinical data, as there is variable-specific information that needs to be considered. This paper proposes an automated data cleaning method for EHR data with clinical knowledge taken into consideration.

Methods: We used EHR data collected from primary care in Flanders, Belgium during 1994-2015. We constructed a Clinical Knowledge Database to store all the variable-specific information that is necessary for data cleaning. We applied Fuzzy search to automatically detect and replace the wrongly spelled units, and performed the unit conversion following the variable-specific conversion formula. Then the numeric values were corrected and outliers were detected considering the clinical knowledge. In total, 52 clinical variables were cleaned, and the percentage of missing values (completeness) and percentage of values within the normal range (correctness) before and after the cleaning process were compared.

Results: All variables were 100% complete before data cleaning. 42 variables had a drop of less than 1% in the percentage of missing values and 9 variables declined by 1-10%. Only 1 variable experienced large decline in completeness (13.36%). Therefore the overall drop in the completeness was acceptable. All variables had more than 50% values within the normal range after cleaning, of which 43 variables had a percentage higher than 70%, meaning that the data quality was largely improved by the cleaning method and the cleaned data was more correct and valid.

Conclusions: We propose a general method for clinical variables, which achieves high automation and is capable to deal with large-scale data. This method largely improved the efficiency to clean the data and removed the technical barriers for non-technical people.

Key words: Data Cleaning, Automated method, Clinical decision support.

2.1 Backgrounds

With the rapid development of electronic systems, the concept of Electronic Health Records (EHR) have already been widely accepted and used, and the use of EHR data in clinical research is incredibly increasing, leading to "a new era of data-based and more precise medical treatment" [73]. EHRs may include a wide range of data, including socio-demographics, medical history, medication, allergies, immunization and laboratory test results. The abundancy of data resources provides sufficient information for clinical studies, but also raises challenge of data cleaning.

The data collected from daily practice may include lots of typos and errors. The data cleaning of real-world data is time-consuming, and requires to master techniques of data processing. Therefore, a tool for non-technical people, which can do the data cleaning automatically, can save a considerable amount of time and budget.

In order to correctly clean the EHR data and evaluate data quality, some studies gave their own definitions of dimensions of Quality of Data (QoD) for EHR data. A systematic review conducted by Weiskopf et al. in 2013 summarized 5 dimensions, i.e. completeness, correctness, concordance, plausibility, and currency [194]. A study in 2017 summarized accuracy, completeness, consistency, credibility, and timeliness [59], while a study in 2019 assessed primary care EHR data considering comparability, completeness, correctness, and currency [178]. The systematic review in 2018 summarized 34 studies with a focus on data quality assessment in emergency medical services [104], classifying the assessment indicators in all these studies into 5 dimensions, namely completeness, accuracy, consistency, accessibility, and timeliness. The sequence of these dimensions is based on the frequency of occurrence in 34 studies.

With the clearly defined dimensions of QoD, some researchers proposed the frameworks to assess the data quality systematically. Kahn et al. (2012)

presented a "fit-for-use" conceptual model for data quality assessment [84], and generated a framework to assess the data and control the data quality. The following studies developed their own frameworks to systematically assess data, by calculating self-defined assessment indicators [164, 118], visualizing variability [56], or adding analysis-specific data quality evaluation requirements [47]. However, these frameworks were only meant to assess the data quality, no data cleaning procedures were conducted.

With the necessity of efficiently cleaning EHR data, several proposals for a general framework were made. A study in 2016 applied the framework proposed by Kahn et al. by defining assessment indicators following their 5 domains and correcting the data if there is room to improve the assessment indicators [49]. Weiskopf et al continued their study in 2013 and published a data quality assessment guideline in 2017 [195], which can help evaluate the data quality from different aspects and hence help solve data issues. The framework presented in the work of Miao et al. functions in a similar way [108]. These methods or frameworks rely on the data quality assessment to accurately identify the potential data issues in the original data, then the researchers worked to solve the data issues accordingly. Manual investigation was throughout the whole process.

Phan et al. (2020) proposed an automated data cleaning method for EHR data [133], but it only considered anthropometric data (height and weight measurements), so that the automated cleaning techniques were based on the data characteristics of height and weight measurements. Tang et al. (2020) developed a pipeline with high automation that can process different types of variables, such as categorical variables, numeric variables, and hierarchical variables [176]. But the main functions of this pipeline is to help extract data, select high-quality features, harmonize data, and merge the data as the final output. Data cleaning for wrongly input values was not discussed.

An automated data cleaning for different types of medical variables is still

lacking. Besides, the current automated data cleaning tools in other areas often use a statistical calculation to process all variables uniformly. However, in clinical practice, there is a lot of variable-specific information that needs to be considered. Therefore, it is necessary to develop an automated method with good clinical relevance. The goal of this paper is to introduce an automated data cleaning method, which was used to clean data collected from Belgian primary care.

2.2 Methods

2.2.1 Study Setting

Intego is a longitudinal database of EHR data collected from patients in general practice from the Flanders region in Belgium. It is a unique operational computerized registration network. The network was founded in 1994 and 111 general practitioners (GP) participated by 2015. Data is yearly collected for every patient who had a contact with his GP in that year. In total, about 300,000 individual patients have been recorded in the database since the network started, corresponding to a yearly coverage of approximately 2% of the Flemish population [184]. In 2015, there were more than 3 million clinical measurements, 3 million diagnosis, 14 million prescriptions, and 38 million laboratory tests of about 10,000 variables.

Table 2.1 shows some example rows of the measurement dataset. The dataset including laboratory tests has a very similar structure. Although Intego also has a prescription dataset and diagnosis dataset, there are no unexpected errors, so they will not be discussed in the paper.

There were over 10,000 laboratory test variables in Intego. It was not possible to clean all variables. The variables used for testing the automated data cleaning method were determined based on feature selection studies that

Table 2.1: An example of raw data

Patient ID	GP ID	Birth Year	Sex	Test Code	Date	Year	Test Name	Num1	Num2	Unit
27898	GP 9811	1957	F	S00000057	41860	2014		165	0	Cm
27898	GP 9811	1957	F	S00000058	41860	2014		56,5	0	Kg
27898	GP 9811	1957	F	S00000057	34971	1995	Height	163	0	S002
27898	GP 9811	1957	F	S00000058	34971	1995	Weight	61	0	S006
42716	GP 9812	1963	M	S00000069	41596	2013	1	130	88	mmHg
61905	GP 9845	1964	F	S00000069	34197	1993	Blood pressure	90	50	mmHg

will be conducted in the future. To avoid bias caused by disease selection, we considered variables related with 3 diseases of different types. We selected and cleaned 52 clinical variables with more than 100,000 observations that were discussed in previous studies about Type 2 Diabetes, myocardial infarction, and rheumatoid arthritis.

2.2.2 Data Descriptions and Issues

The main data issues for data cleaning are situated in the values and units in the records of measurements and laboratory tests.

Values out of normal range

There are two types of value errors. The first type of errors was caused by the fact that more or less zeros than expected were entered. Take blood pressure as an example, the wrong values can be 9, which should be corrected into 90, or 16000, which should be corrected into 160. This type of errors can be easily converted back to the normal range by a simple transformation in the order of magnitude. The second type of errors is the unreasonable values that cannot be converted back to the normal range using simple transformation. For instance, 3000 for blood pressure is an unreasonable value even after converted as 30 or 300. This kind of errors should be detected as outliers and replaced with NA. Therefore, it is necessary to

know the normal range and outlier range for each variable if we would like to clean the data following real clinical practice.

Inconsistent units

There are also two reasons leading to inconsistent units. First, the lab tests can come from different labs across the country, and the labs and GPs are not using uniform units for the same variable. For example, the clinical variable Creatinine has units such as mg/dl, mmol/L, mg%, which are all correct units but should be converted into a uniform format. Second, some of the units were wrongly spelled, e.g. ‘g/dL and mmol/mol, or meaningless, e.g. 77, 1, NULL, etc. In order to have uniform units, the wrongly spelled units should be automatically detected and replaced with correct units, and then all the units of one variable are converted into one uniform unit. One difficulty of unit conversion is that the conversion rates between the same units for different variables are different. For instance, the conversion rate between mmol/L and mg/dL for glucose is 18 while it is 11.312 for creatinine. Thus, the variable-specific information should be known in advance for better data cleaning.

2.2.3 Clinical Knowledge Database (CKD)

The Clinical Knowledge Database (CKD) stores all the data related with variable-specific rules that is necessary for data cleaning. In our use case, we included the conversion formula between different units, and the normal range and extreme range of the clinical variables, however, it is not necessary to keep CKD in a uniform format all the time. It can be easily adapted to other studies if more variable-specific information is needed. For example, it can store the variable-specific statistical measures to aggregate the repeated measure.

The unit conversion rates between all possible units for a certain clinical variable were automatically derived from the website http://unitslab.com/ without manual selection. The standard unit, the normal range, and the extreme range of clinical variables were chosen based on literature and clinical knowledge. The normal range is the reference range for the healthy people and the extreme range is the biologically plausible range. The clinical information can be easily filled in by clinicians after this automated cleaning tool is put into use. Table 2.2 gives an example of CKD and the full CKD is uploaded online (https://github.com/XiShi6/Automated-data-cleaning).

Table 2.2: Example rows in Clinical Knowledge Database

VariableCode	VariableName	VariableUnit	ConversionRate	StandardUnit	Normal Range		Extreme Range	
					Max	Min	Max	Min
57139A.B	Creatinine	Mg/dL	1	Mg/dL	0.6	1.3	0.1	5
		Mmol/L	11.312	Mg/dL				
		Mg/L	0.1	Mg/dL				

2.2.4 Fuzzy Search

In order to correct the typos in the units, the fuzzy matching algorithm based on Levenshtein distance was applied to automatically detect the wrongly spelled units and replace them with correct units or NA if no similar units could be found. Levenshtein distance is the minimum edit distance that contains a set of editing operations, including changing characters, deleting characters and/or inserting characters. When the words are identical, the distance is zero. When one letter has to be changed/deleted/inserted, the distance becomes one. Therefore, three types of errors are defined, i.e., deletion, insertion, and substitution. Denote V and W as two words containing letters $i=1,\ldots,n$ and $j=1,\ldots,m$, and ϵ is the empty word, the Levenshtein distance between these words for different types of errors can

be defined as the equations below [154].

$$Lev(V, \epsilon) = |V| = n \tag{2.1}$$

$$Lev(\epsilon, W) = |W| = m \tag{2.2}$$

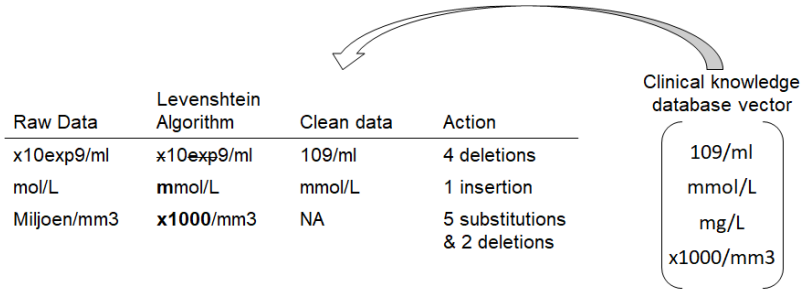
$$Lev_{V,W}(i,j) = \begin{cases} max(i,j) & \text{i=j} \\ min \begin{cases} 1.Deletion : Lev(i-1,j)+1 \\ 2.Insertion : Lev(i,j-1)+1 \\ 3.Substition : Lev(i-1,j-1)+1 \end{cases} & \text{i} \neq \text{j} \end{cases} \tag{2.3}$$

The Levenshtein distance has the disadvantage when comparing words at different length, since longer words are more likely to have larger distance. A similarity ratio can be calculated from Levenshtein distance to standardize the distance and make it independent of the word length [151]. The obtained ratio ranges between 0 and 1, with 0 meaning no similarities and 1 meaning identical. The equation for the similarity ratio can be written as below, where Lev is the Levenshtein distance, $|V|$ and $|W|$ are the number of characters. Table 2.3 gives an example of the similarity ratios between some units.

$$Lev.ratio(V,W) = 1 - \frac{Lev(V,W)}{max(|V|,|W|)} \tag{2.4}$$

Table 2.3: An example of the similarity ratio for the unit check

Original Unit	Correct Unit	Score
106/L	X10exp6/L	0.71
mmol/LsOms	mmol/L	0.75
g/24u14g/L15	g/24u	0.61
77	g/24u	0
24	g/24u	0.57



Raw Data	Levenshtein Algorithm	Clean data	Action
x10exp9/ml	x10~~exp~~9/ml	109/ml	4 deletions
mol/L	**m**mol/L	mmol/L	1 insertion
Miljoen/mm3	**x1000**/mm3	NA	5 substitutions & 2 deletions

Figure 2.1: A toy example to show how the fuzzy search method works. The correct possible units for one specific variable are known in CKD. The algorithm compares the raw data and the possible units in CKD and selects one unit that requires a minimum of edits or NA if the scores are below the threshold. (The vector of correct unit is not the full vector in CKD.)

Figure 2.1 gives a toy example how the fuzzy search works in the real dataset. All possible units for one specific variable are stored in the column *Variable Unit* in CKD. The algorithm computes the scores between the original unit and every possible unit of this variable and replaces the original unit with the correct unit, which has the largest similarity ratio. If the original unit is correctly spelled, there will be one identical unit in CKD, then the original unit will be replaced with this identical unit, i.e., keeps unchanged. We conducted the replacement with the identical unit in order to process all the units in a uniform way. Next, for incorrectly spelled units the similarity ratio is set at a specific threshold. If the similarity ratio is smaller than the threshold, the original unit is replaced with NA. In order to select the optimal threshold, we did a grid search by the step of 0.1 to find the threshold leading to the minimum error. After examining the results from Fuzzy search and compared the results of the grid search, we defined the threshold as 0.5 as it gave the most valid and reasonable result.

2.2.5 Outlier detection

Traditional outlier detection methods that depend on the data distribution to automatically detect values strange or far from the majority of the values could not perform well in our use case. Intego is representative of the general population, so both healthy people and patients are included. If the variables have a large range in values, healthy people and patients can have very different observations. As a consequence, traditional outlier detection methods will detect patients' records as outliers if the disease prevalence is low, which is almost always the case in primary care.

Van den Broeck et al. (2005) proposed an outlier detection method based on variable-specific information [44]. They specified the true normal, true extreme, erroneous and idiopathic values for each variable and discussed the outliers case by case. Their method could not easily be applied in a big data and automated framework, thus, it was simplified into an outlier detection method based on the normal range and extreme range of variables.

If the values were out of the normal range, a transformation in the order of magnitude was first conducted, trying to make the values as close to the normal range as possible, e.g. 9 for blood pressure could be converted into 90 (Type 1 error in Section 2.2.2). If the values were still out of the extreme range after the transformation, they were regarded as outliers and replaced with NA (Type 2 error in Section 2.2.2), e.g. 3000 for blood pressure was replaced with NA because it is an unreasonable value even when it was converted as 30 or 300. Figure 2.2 shows an example of the outlier detection by using the data range of clinical variables and the algorithm is shown as below.

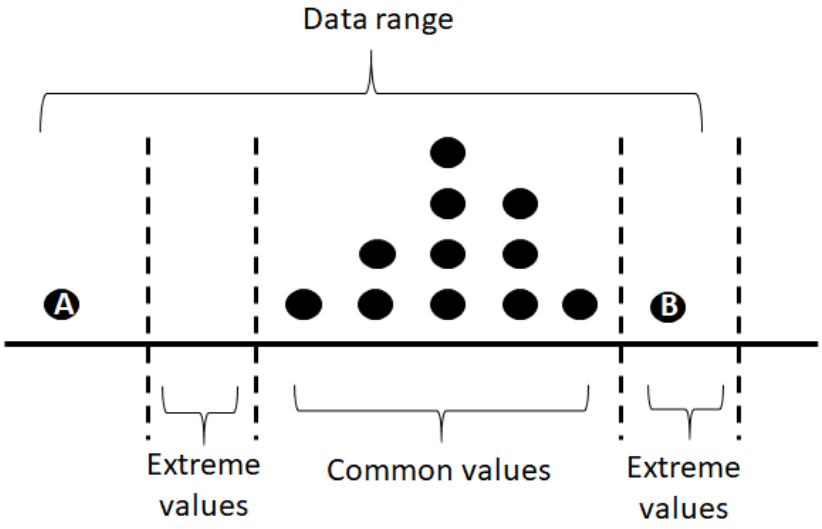


Figure 2.2: Example of a data range. It is assumed that the majority of the data falls within the common range. All data outside the extreme range are outliers. This makes data point A an outlier and data point B a more extreme but acceptable observation.

Algorithm: Outlier correction algorithm

If $result < 0$ AND $Min\_Common >= 0$:
    $result = -1 * result$
If $result == 0$ AND $Min\_Common > 0$:
    $result = NA$
While ($result > Max\_Common$ AND $result/10 > Min\_Common$) OR ($result > Max\_Extreme$):
    $result = result/10$
While ($result < Min\_Common$ AND $result * 10 < Max\_Common$) OR ($result > Min\_Extreme$):
    $result = result * 10$
If $resultnotin(Min\_Extreme, Max\_Extreme)$:
    $result = NA$

2.2.6 The whole cleaning framework

With the techniques mentioned above, an automated data cleaning method was designed which could clean the data following the real clinical practice (Figure 2.3).

The records of one specific variable were extracted from the whole database by using the *Lab Code* in CKD, then the preprocessing procedure was conducted to make a preliminary preparation, such as removing unexpected characters from the values, converting the values as numeric format and so on. The follow-up data cleaning procedures, i.e. fuzzy search to correct the wrongly spelled units, unit conversion, and outlier detection, were done as described in previous sections. Afterwards the cleaned variable can be stored in the data warehouse for follow-up studies.

The Clinical Knowledge Database, the codes for the automated data cleaning method, and the complete list of completeness, correctness, and plausibility can be found in Section Availability of data and materials

2.2.7 Evaluation measures

There is no consensus how to evaluate data quality, especially for EHR data, but some studies gave their own definitions of dimensions of Quality of Data (QoD) for EHR data. A systematic review in 2013 summarized 5 dimensions, i.e. completeness, correctness, concordance, plausibility, and currency [194]. Completeness is the percentage of missing data, and correctness is the percentage of correct and accurate data. Concordance means comparability and consistency between data elements, for example, the prescriptions taken by the patients are in correspondence with the diagnosis. Plausibility means validity and credibility, which can be tested by checking the percentage of values falling within the biologically plausible range, i.e. the extreme values in our paper. Currency evaluates the timeliness and regency of the

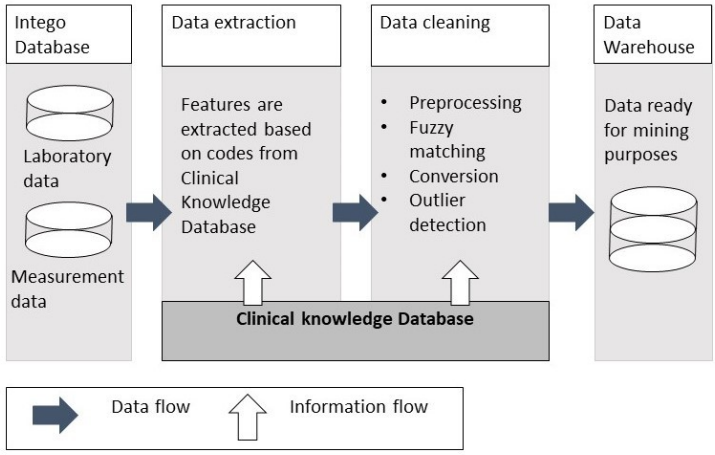


Figure 2.3: The data cleaning framework to build a data warehouse to be used for follow-up studies.

data. A study in 2017 summarized the quality dimensions as accuracy, completeness, consistency, credibility, and timeliness [59], while a study in 2019 proposed a model to assess primary care EHR data, covering comparability, completeness, correctness, and currency [178].

The concordance/comparability/consistency and currency/timeliness of data is not affected by our data cleaning method, so they are not evaluated in the paper. In order to evaluate other dimensions, we calculated the percentage of missing values for completeness, and the percentage of values within the normal range for correctness. Any observations outside the extreme ranges have been detected as outliers and replaced with NA, therefore, all the values are biologically plausible and plausibility is always guaranteed.

2.3 Results

The flowchart of the whole cleaning method is shown in Figure 2.4, using Erythrocyte as an example to give an idea how the variables were affected in each step. Table 2.4 shows the percentage of missing values and the percentage of values within the normal range before and after applying the automated data cleaning method. Only the top 10 variables with the largest volume are shown in the table, the full table has been uploaded online. There were no empty cells in the column for numeric values in the original data because of database maintenance, so the initial missing rates for all variables were 0. However, there were some cells with meaningless information, e.g. “,”, “n”, “NULL”, “na”, which were detected and converted into NA in the first step, data preprocessing. The completeness was not improved by using our method, because we didn’t impute missing values to increase completeness, on the contrary, it slightly decreased because we replaced the outliers with NA. All variables were 100% complete before data cleaning. In total, 42 variables had a drop less than 1% in the percentage of missing values and 9 variables declined by 1-10%. Only 1 variable experienced large decline in completeness (13.36%).

However, a slight loss in completeness is acceptable because the correctness was largely improved. Based on the percentage of the values within normal range before and after the cleaning, the correctness of most variables were largely improved, for example, Monocyte from 14.12% to 69.93% and Leukocyte from 34.04% to 98.21%. All variables had more than 50% values within the normal range after cleaning, of which 43 variables had a percentage higher than 70%. The level of data quality in the raw data varied seriously for different variables. Some variables had good quality even before the data cleaning, such as Hemoglobin (92.68%), Hematocrit (94.02%), and MCV (92.53%). In this case, our method could not further improve the data quality by a large margin, but still lead to slight improvement, e.g. from 94.02% to 97.17% and from 92.53% to 92.8%. The correctness of 9

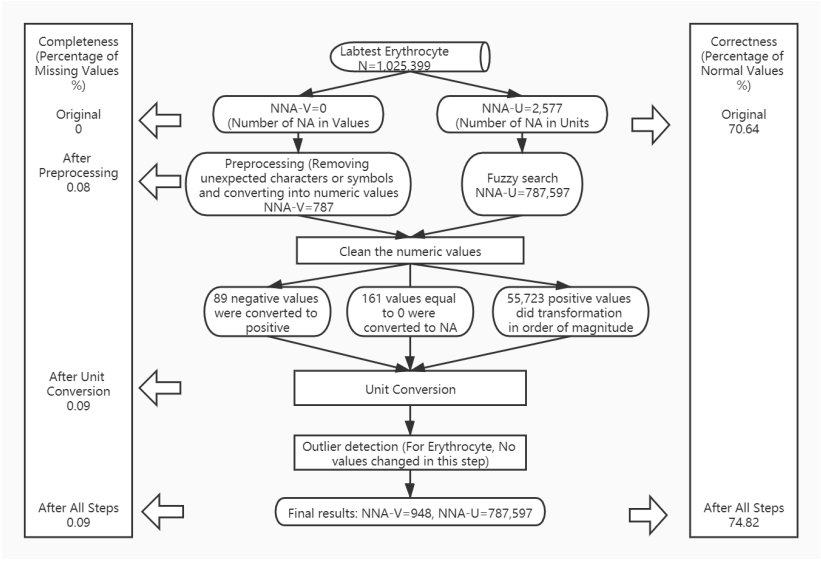


Figure 2.4: The flowchart of Erythrocyte to give an example of the changes in each step

variables were still lower than 70% after data cleaning, of which 6 variables had an increase smaller than 5% because of the cleaning procedure.

As mentioned in Section 2.2.7, the measure that is used for plausibility is often the percentage of values within extreme range. Values outside extreme ranges were detected as outliers and replaced with NA, meaning that all the values were biologically plausible. The plausibility before data cleaning is in the online document (https://github.com/XiShi6/Automated-data-cleaning)

2.4 Discussion

We proposed an automated data cleaning method for clinical variables. One of the main contributions of this method is that it allows non-technical

Table 2.4: Completeness and correctness before and after data cleaning (Top 10 variables with the largest volume)

Test Name	Completeness				Correctness		Number of Observations
	Original	After preprocessing	After unit change	After all steps	Original	After all steps	
Hemoglobin	0	0.02	0.02	0.03	92.68	92.71	1061333
Lymphocyte	0	0.03	0.03	0.04	14.23	54.1	1060664
Eosinophils	0	0.03	0.03	3.87	15.71	96.43	1055109
Monocyte	0	0.03	0.03	0.26	14.12	69.93	1053768
Basophil	0	0.06	0.06	0.06	17.91	73.09	1027615
Hematocrit	0	0.02	0.02	0.02	94.02	97.17	1025484
Erythrocyte	0	0.08	0.08	0.09	70.61	74.82	1025399
Leukocyte	0	0.1	0.1	9.7	34.04	98.21	1013012
MCV	0	0.03	0.03	0.24	92.53	92.8	1003326
MCH	0	0.03	0.03	0.24	84.16	84.42	997867

Note: Completeness is Percentage of Missing Values (%) and Correctness is Percentage of Normal Values (%).

people to explore real-world data without needing proficient data processing techniques. Although there is abundant EHR data stored in primary care, hospitals and laboratories, the technical barrier of data cleaning leads to the fact that only a small group of people, such as statisticians and data analysts, can make full use of EHR data. Take Intego Database as an example, there are 38 million laboratory tests of about 10,000 variables. In general, it may take one proficient data scientist hours to clean one variable. Then it is almost impossible to clean all the data within a reasonable period. A massive amount of data remains untouched. However, it can also provide insights and supports for decision-making if a broader range of people, e.g. clinicians and health policy makers, can freely get access to the clean and organized data in a convenient way. Despite the high automation is at the cost of slight decline in completeness, it is still reliable to get results of general trends and associations.

Another contribution of this method is it largely improved time efficiency with a slight loss in accuracy. It can take a proficient data scientist 30-40 hours to clean around 50 variables manually. The time of manual data cleaning was estimated based on the past experience when the laboratory or

measurement variables were cleaned manually for previous studies [138, 61]. It took 5.2 minutes by using the automated data cleaning method. It will be more convincing if a group of data scientists can clean the same variables manually and compare the average time with the automated method. However, it was not possible to conduct such experiment in this study because of lack of resources. The time to construct CKD was not counted here. Currently the CKD was constructed by the authors whose background is machine learning and the reference data was found in literature, taking a lot of time to find extreme values. However, when the real users, i.e. the clinicians, use the tool and update the CKD, they can easily enter the reference data based on their clinical knowledge and clinical practice, meaning that it only takes seconds to do so.

As mentioned in Background, there are some data preprocessing software or tools that can deal with commonly observed data issues. For example, the R package dataQualityR can perform variable-level data quality checks and generates summary report [3]. But these software and tools share one common characteristic, they detect abnormal values based on the data distribution and clean the data by using some summary statistics. However, with the real-world data, when there are a considerable amount of original values as abnormal values, the data distribution already departs from the distribution of the true values. Moreover, there are clinical considerations for clinical variables when cleaning data, which is from expert knowledge and is not included in data itself. Thus, most of the data preprocessing tools cannot perform well for clinical variables automatically. Manual investigation is needed in each step to make decisions based on the data reports. Our automated data cleaning method have better clinical relevance by constructing the Clinical Knowledge Database (CKD). We cleaned the numeric values not based on the data distribution, but the objective clinical knowledge, such as normal range and extreme range of clinical variables. The use of CKD also made it possible to do unit conversion on the variable level following the clinical practice. For instance, the conversion rate between

mmol/L and mg/dL for glucose is 18 while it is 11.312 for creatinine.

In fact, the current study is the initial exploration of using CKD to achieve automated data cleaning for EHR data. Therefore only the most basic cleaning procedures were included in this paper. It is possible to further expand the CKD in the future by adding more clinical knowledge. For instance, the statistical measures for each variable can be added in case of data aggregation. Mean, median, min, max, etc., different measures can be used for different variables based on specific clinical considerations. It is possible to go one step further, adding more complicated "IF-THEN" rules into CKD, to help improve consistency. For example, absolute band cell count and relative band cell count are similar concepts, the difference is that one is absolute count number and the other is the percentage. If the value of relative count (a percentage) is mistakenly entered as an absolute count, it will be detected as an outlier and replaced with NA. By adding one consistency rule in CKD, the algorithm can automatically check whether this is a percentage number with unit measurement "%", meaning that it has been assigned a wrong variable code (See variable code in Table 2). These examples reveal the potential of CKD to help achieve better clinical relevance and consistency in the future studies. The CKD acts as an expert or dictionary and the automated data cleaning method proposed in this paper guarantees that the information in CKD is made the biggest use of without losing high level of automation.

Apart from the use of CKD, we applied the fuzzy search method to automatically clean the measurement units, which, to our knowledge, is novel for cleaning medical text data. This method has not been used in previous studies with a focus on data cleaning of EHR data. Our results have shown that it could perform well for medical text data and has a great potential to be used for unstructured data cleaning in the future. Another methodological contribution is the outlier detection method. An outlier detection method was proposed in 2005 [19], which diagnosed all

the data points based on the true nature of the data, using hard and soft cut-off values. We adjusted this method to make it suitable for big data, and successfully incorporated it with CKD, which helped to achieve an automated process for medical data of different types.

Based on the literature review in Background, although there were some studies using data assessment frameworks to help identify potential data issues in the original data, hence help clean the data, the level of automation was relatively low in such studies. The assessment frameworks could systematically examine the data quality, but the decision of data cleaning still needs to be done by researchers. Manual investigation cannot be avoided in the whole process. To our knowledge, two studies (Phan et al. 2020 and Tang et al. 2020) proposed an automated data cleaning method or pipeline, but both studies have inherent limitations. Phan et al. (2020) only included height and weight variables, while Tang et al. focused on data preprocessing tasks such as data extraction and data harmonization, rather than data cleaning. Our method achieved to efficiently clean different types of clinical variables, with clinical relevance taken into consideration.

One limitation of this method is that data longitudinality and patient variability are not considered, making it unable to check and improve data consistency. By considering the longitudinal data from the same patients, the correctness of some variables, which have large gaps between normal range and extreme range such as total cholesterol, may be further improved. However, the current method does not have auto-correction based on the data distribution, because when the correction is done automatically following some kind of distribution without the actual investigation of data, it can be very risky considering the automated method is applied for all kinds of clinical variables. Imputation bias may be introduced because of the auto-correction. In the current method, all the changes in the values are done based on the "clinical information", which are the objective facts independent from the database. So the cleaning procedure will not

be affected by the errors in the original values, meaning that this is the basic procedure that anyone will do in the same way. Therefore, the auto-correction based on longitudinal information is not considered in the current method to guarantee validity and credibility.

To overcome the limitation on the basis of the current method, it is reliable to add longitudinal information if it is done following some pre-defined rules instead of data distribution. As mentioned above, it is possible to add more complicated consistency rules in CKD to help improve the performance of the method, similarly, the rules about longitudinal characteristics of each variable can also be added in CKD. Here is one example to demonstrate how CKD can help improve the longitudinal outlier detection in the future studies. The normal level of cholesterol is below 200 mg/dL and the extreme maximum value is 1000 mg/dL. A patient may have a series of records, 189, 174, 700, showing that the patient had healthy records at the beginning but probably got health issue when the last record occurred. There is a risk that the record of 700 is identified as "outlier" if the auto-corrected outlier detection is conducted based on longitudinal data. In this case, we exempt such values using the CKD information as they are within the extreme values defined in CKD. Another example is the possibility of adding longitudinal rules in CKD. If a patient has a record of 702, 890, 970, 1030, the last record "1030" should be replaced with NA as it is out of the extreme range. However, it is likely that this is a patient with severe symptoms. To avoid this type of issue, a longitudinal rule may be added for this specific variable, such as "IF the former values are valid and the follow-up extreme value is closely to the former values within a certain distance THEN do not replace the extreme value with NA". The choice of "certain distance" may be variable-specific information. In summary, defining rules to detect abnormal values in the longitudinal data requires extensive discussion with data and clinical experts, which is expected to be done in future studies.

2.5 Conclusion

While some studies proposed a general framework to clean and assess EHR or other formats of medical data, manual investigation was still needed in the whole process, very limited research studied an automated data cleaning method or tool for EHR data. Moreover, a general automated data cleaning method or tool that is able to clean all kinds of clinical variables is lacking. Furthermore, the automated data cleaning tools in other domains are not capable to accurately clean clinical data with variable-specific information taken into consideration.

This paper proposed a method which could clean clinical data automatically and is capable to work on large-scale data. Primary care EHR data was used to test the method. Results showed that with a slight loss in completeness, the credibility, correctness, and plausibility of the data were largely improved after the data cleaning. The method can be used and flexibly adjusted to clean EHR data in other studies in the future.

Chapter 3

A Feature Selection Framework

This chapter has been published by BMC Medical Informatics and Decision Making as:

Shi X, Nikolic G, Epelde G, Arrue M, Bidaurrazaga Van-Dierdonck J,Bilbao R, De Moor B. (2021). An Ensemble-Based Feature Selection Framework to Select Risk Factors of Childhood Obesity for Policy Decision Making. BMC Medical Informatics and Decision Making:21:222.https://doi.org/10.1186/s12911-021-01580-0

ABSTRACT

Background: The increasing prevalence of childhood obesity makes it essential to study the risk factors with a sample representative of the population covering more health topics for better preventive policies and interventions. It is aimed to develop an ensemble feature selection framework

for large-scale data to identify risk factors of childhood obesity with good interpretability and clinical relevance.

Methods: We analyzed the data collected from 426,813 children under 18 during 2000-2019. A BMI above the 90th percentile for the children of the same age and gender was defined as overweight. An ensemble feature selection framework, Bagging-based Feature Selection framework integrating MapReduce (BFSMR), was proposed to identify risk factors. The framework comprises 5 models (filter with mutual information/SVM-RFE/Lasso/Ridge/Random Forest) from filter, wrapper, and embedded feature selection methods. Each feature selection model identified 10 variables based on variable importance. Considering accuracy, F-score, and model characteristics, the models were classified into 3 levels with different weights: Lasso/Ridge, Filter/SVM-RFE, and Random Forest. The voting strategy was applied to aggregate the selected features, with both feature weights and model weights taken into consideration. We compared our voting strategy with another two for selecting top-ranked features in terms of 6 dimensions of interpretability.

Results: Our method performed the best to select the features with good interpretability and clinical relevance. The top 10 features selected by BFSMR are age, sex, birth year, breastfeeding type, smoking habit and diet-related knowledge of both children and mothers, exercise, and Mother' s systolic blood pressure.

Conclusion: Our framework provides a solution for identifying a diverse and interpretable feature set without model bias from large-scale data, which can help identify risk factors of childhood obesity and potentially some other diseases for future interventions or policies.

Keywords: Feature Selection; Ensemble Learning; Childhood Obesity; Public Health; Policy decision making.

3.1 Background

Childhood obesity has emerged as an important public health problem all around the world. According to the WHO [123], the worldwide prevalence of obesity nearly doubled during 1980-2008 and one in three 11-year-old children is overweight or obese in Europe. A childhood obesity review has shown that the increasing prevalence of childhood obesity is associated with the emergence of comorbidities previously considered as "adult" diseases such as diabetes and hypertension which can track into adulthood [92]. Basque Government proposes the creation of specific childhood obesity prevention plan as one of the new two main action areas and targets to create regional policy for the Basque Region [79]. Therefore, it is essential to study the risk factors of childhood obesity to design preventive policies or interventions.

The increasing prevalence of childhood obesity is the consequence of an interaction among a complex set of factors that are related to the genetics (e.g. monogenic disorders, endocrine issues, etc.) and environmental ecological effects (e.g. parental feeding styles, microbiota, school, society trend, etc.) [92]. There are some surveys capturing the reality of overweight and obese children, however, most of the surveys and studies only included hundreds or thousands of participants. Based on the summary statistics from the systematic reviews [32, 100, 99], very limited studies had more than 100,000 or even over one million participants. The rapid development of electronic health records (EHR) systems and abundant data from various sources make it possible to have access to real-world data on a larger scale. A review in 2018 explored the obesity studies using big data collected from different sources [180], such as social media, smartphones and healthcare wearable devices, transportation and so on. But these data samples had their own limitations, for example sample bias, ethical issues, or lack of linkage with nutrition information. A recent study used EHR data to predict the risk of childhood obesity, including almost one million participants [125],

however, because of the characteristics of data sources, the features were all clinical variables, the environmental factors related with family and school were not included. Therefore, the Osakidetza database in Basque region can be of great value as it is large scale data covering millions of participants and includes specific information on different aspects of environmental factors of childhood obesity at the same time.

Studies on the risk factors of childhood obesity and interventions have been conducted and reported in the literature [92, 32, 100]. But difficulties still exist when designing preventive policies or interventions for a specific region. Despite that the risk factors given by different studies were similar, the variables used for each risk factor could be very different. For example, it is known that parental feeding styles have a remarkable influence on childhood obesity. Some studies drew the conclusion based on a questionnaire on feeding style, some used one categorical variable to summarize the general type, while some used calorie-intake or sugar-intake. Thus, if a region would like to design the customized policy or intervention that fits the local environment the best, it is difficult to make the decision which indicator can represent a general aspect most accurately, then it is unclear how to design corresponding interventions. In this case, applying machine learning models with local comprehensive data can help identify the most influencing variables grounded in real-world data, which can reduce the cost of testing factors and accelerate the establishment of policies or interventions that are more "evidence-based". In addition to assisting on the localization of scientific research derived knowledge-based policy making, machine learning techniques can help confirming, with a wider real-world data, the results of scientific rationale and research for a specific region or timeframe, and occasionally derive insights or new hypothesis to be tested by scientific research and be candidate for adoption by policy decision making (if it's validated and relevant). Therefore, it is expected that machine learning models can contribute to further scientific research as wells as to provide supports for policy decision making, providing domain experts with an

additional tool to support their decisions making processes.

To reduce the features of the data, two main classes of machine learning models can be applied, namely feature extraction methods and feature selection methods. The difference between the two types of models is that feature selection methods keep a subset of the original variables while feature extraction methods combine the original variables into a smaller set of new features. In our study, we focus on feature selection methods to preserve the semantics of the features, as the results need to be interpretable without a subjective definition of new features. Moreover, it is easier for clinicians and policy-makers to establish follow-up interventions for a single feature than a compound factor of multiple features.

There are three main types of feature selection methods: filter methods, wrapper methods, and embedded methods [13]. Filter methods select features based on a statistical measure to assign a score to each variable to rank the variable importance regardless of the model. These methods are time-efficient and robust to overfitting, but tend to ignore the possible interactions between variables. Wrapper methods have the opposite advantages and disadvantages by converting the feature selection task into a search problem. Subsets of variables are compared with other subsets to select the group of features that can give the best predictive performance. Embedded methods are learning models that can perform feature selection and classification simultaneously by integrating feature selection algorithm as part of the learning process. The embedded methods take variable interactions into consideration and are less computationally demanding than wrapper methods. However, in some cases, the optimal feature set selected by one embedded method is classifier-dependent, meaning that the optimal set only works for this specific classifier and cannot contribute to good prediction when used for other embedded classifiers [74], as the optimal set is based on the hypotheses of the classifier.

Different feature selection methods have their specific advantages and

disadvantages, resulting in the discrepancy in the selected variables. Hence, the results may be biased because of the model limitations if we only rely on one method. This problem is even more critical in exploratory research when the problem is not clear and the validity and credibility of the results are crucial. In addition, most feature selection methods, such as wrapper methods and embedded methods, select features based on predictive performance, which may lead to a selection of variables with no clinical relevance and interpretability. Finally, scalability of the model, i.e. the capability of handling a growing amount of data, is very essential when using a large-scale sample representative of the population. This means special attention is required in terms of time efficiency, data storage, data loading, and so on. Especially in our use case, it was not possible to load the whole dataset in memory at once, so it was an essential demand to develop a model with good scalability.

To overcome these limitations and solve the problems in a real-world setting, we propose BFSMR (the Bagging-based Feature Selection framework integrating MapReduce), a machine learning method that can perform efficiently with large-scale data and combine the results from different feature selection methods to give a more convincing and interpretable selection of features. A higher clinical relevance and a better interpretability of the algorithm-identified risk factors can give suggestions that are easy to implement in practice, which helps to better shape policies and corresponding interventions to overcome a specific public health challenge. Moreover, when using multiple feature selection methods as components of BFSMR, we can get a comprehensive understanding of the potential important risk factors preferred by different models, meaning that we are providing comprehensive information that is helpful for the decision-making process.

3.2 Methods

3.2.1 Participants and data source

This study uses data from the Osakidetza databases of the Public Health Provider in the Basque Country (Spain) [124], which provides services to more than 2,200,000 patients, through 16 hospitals and more than 300 primary health centers. This is a database recording information about patients for a global view of health status, not specifically for childhood obesity. The Osakidetza database was properly anonymized and extracted with the approval of an ethic committee in the Basque Country.

The database contains the information from around 400,000 children who were under 18 years old during 2000-2019 throughout the Basque Country and includes information of these children until they turned 18 years old. Three datasets from the database have been used for this study: children' s information table, children' s forms and the children' s mothers' forms. Within Osakidetza, forms are health and lifestyle questionnaires that the GP or the nurse may or may not fill out in a consultation, recording the status of patients at that time point. Questions in these forms cover different aspects of lifestyles and health status, including physical exercise, smoking habits, drinking habits, diet styles, vegetable consumption, breastfeeding, gestational information, perinatal information, disease history and family disease history, etc.

The original information was stored in 5 separate datasets, the cleansing and merging process is explained in detail in supplementary material Section 3.6.1 and the final merged data has 1,478,857 records from 426,813 children (Selection and exclusion criteria are shown in Figure 3.1). The multiple records from the same child disperse in an age span of 1-18 years old, representing different stages of childhood. Frequent repeated measures during a short period were rarely observed and the experiment using mixed

models showed that the longitudinal effects were not significant, therefore the personal-level longitudinality was not taken into consideration in the algorithm, instead, all the records were regarded as separate records. Besides, there were in total 119 variables used for feature selection, including numeric variables and dummy variables generated from categorical variables. Data pre-processing is detailed in Section 3.6.1. Due to the fact that in children' s forms different input variables are requested depending on the primary care visit motivation, missing rates varied based on the types of variables. The numeric variables about basic health status, such as blood pressure, age, mother' s BMI, child' s birth height and weight, etc., had very few missing values. While other numeric variables focusing on disease risks or life styles had much more missing values, as these questions were asked only when they were applicable. The missing rates for these variables varied between 30% and 70%. Most of the categorical variables also had a missing rate within this range, with very limited categorical variables having extremely high missing rate (over 80%), such as Diet Intent Change and Breastfeeding Abandonment.

3.2.2 Outcome indicator

The indicator of childhood obesity was created based on the age- and gender-specific BMI. The European Childhood Obesity Group defines it as obesity if the BMI is higher than the 97th percentile of the age- and gender-specific subgroup in the reference population. A BMI higher than the 90th percentile is defined as overweight [136]. This study is aimed to observe both overweight and obese state and the 90th percentile is included in our statistics [51], so our outcome is to indicate whether the child has a BMI higher than the 90th percentile of the reference group.

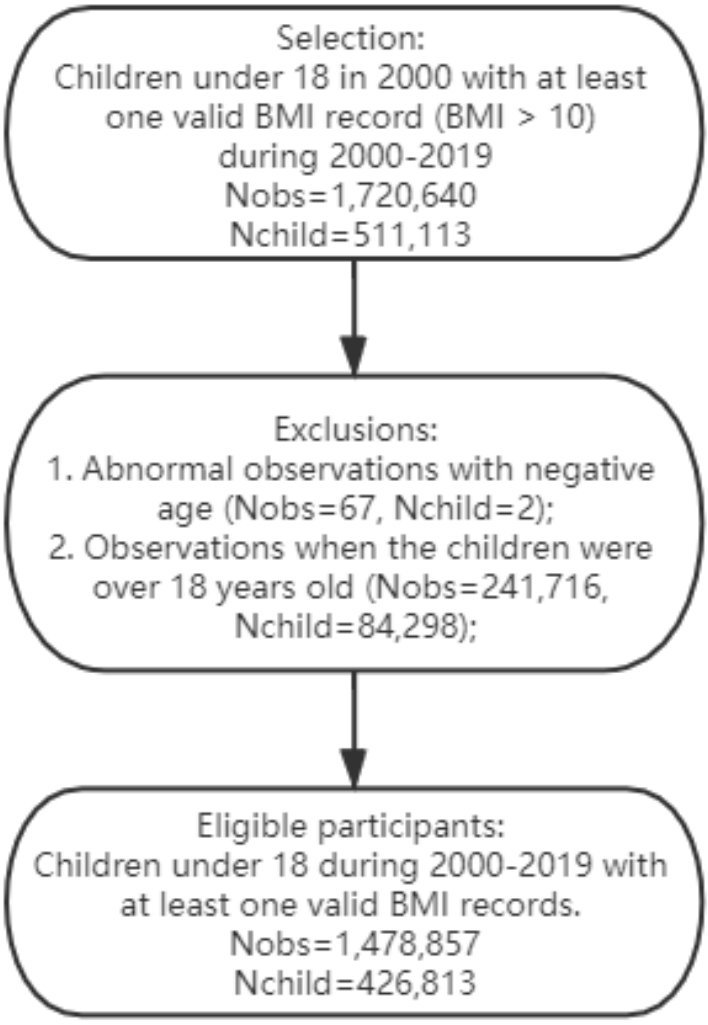


Figure 3.1: The selection and exclusion criteria for the participants.

3.2.3 Embedded feature selection methods

The BFSMR is a bagging-based feature selection framework integrating MapReduce, which is a method with a balance between valid results and good interpretation. In this section, we first introduce the MapReduce technique, the bagging method, and the feature selection models used in the bagging framework. Then we construct a new framework incorporating the advantages of MapReduce and bagging at the same time.

MapReduce

MapReduce is a method to process and generate large-scale data in a parallel and distributed way, which can be very useful in the context of feature selection when the dataset is large-scale and high-dimensional. After splitting the input data into smaller subsets, the model extracts the information of interest in each subset and then merges them to output the aggregated results, ensuring to process large-scale data rapidly. To make it easier to understand, this can be summarized as "split-apply-combine" strategy.

The whole procedure can be broken down into two main tasks, Map and Reduce [42]. The original data is split into an appropriate size and each split is assigned with one Map function defined with respect to data structured in (key, value) pairs. The Map function works in parallel to convert every pair in the input data, denoted as $(k1, v1)$, into a list of pairs in a different data domain, denoted as $(k2, v2)$. Next, all the pairs $(k2, v2)$ with the same key are collected to form one group for one key. Then the Reduce function is applied to each group in parallel and the collection of all Reduce calls is the final result. The procedures of splitting and mapping makes it possible to process the data in parallel and the procedures of shuffling and reducing

merge the information by key variable to reduce the data scale.

$$Map(k1, v1) \rightarrow list(k2, v2) \tag{3.1}$$

$$Reduce(k2, list(v2)) \rightarrow list((k3, v3)) \tag{3.2}$$

Bootstrap Aggregating (Bagging)

Bootstrap aggregating, also called bagging, is an ensemble-learning algorithm that applies different models with different random samples and uses majority voting to combine results for the final decision [27]. The method is incorporated in our model to merge results from different feature selection methods. Given a training data D of size N with correct labels $\omega_l = \omega_1, ..., \omega_C$ representing C classes, generate T bootstrapped samples D_t of size n by random sampling from D uniformly and with replacement. The model m can be applied with D_t to construct the classifier h_t. With the ensemble $\varepsilon = h_1, ..., h_T$ derived from the training process, the unlabeled instance in the testing data is classified into the class that receives the highest total vote.

Feature selection methods

In this study, we selected five models as representative methods from filter, wrapper and embedded methods to give a relatively comprehensive discussion and comparison of the feature selection methods. However, it is not necessary to always use the same feature selection methods. The applied methods can be determined based on the concrete problem and data issues.

We chose Mutual Information (MI) as the statistical measure for filter methods [91], and applied SVM-RFE as one type of wrapper methods [67]. In addition, we used 3 different models from embedded methods, namely

Lasso Regression [179], Ridge Regression [75], and Random Forest [28]. Multiple embedded methods were used, because embedded methods have very different characteristics, making it difficult to use one to represent the diversity of embedded methods. We selected these methods because they are representative methods of filter, wrapper, and embedded method, and they are often discussed in other feature selection studies.

BFSMR

With all the models and techniques introduced above, we propose a framework that combines the advantage of MapReduce and Bagging and gives a more reasonable set of selected features with better interpretability. The notations used in this section are listed as below.

Notation	Meaning
$D = d_p$	D is the input data and is split into P chunks, denoted as $d_p, p = 1, ..., P$
c_i	Feature selection classifier where $i = 1, ..., M$
cid	The Classifier ID where $cid \in 1, ..., M$
s	Random sample set with Set ID $sid \in 0, 1, ..., M$ (sid=0 for test set)
$w1_j$	Feature weights based on the ranking from each classifier where $j = 1, ..., k$
$w2_i$	Method weights based on the model performance where $i = 1, ..., M$
f_i	Feature lists derived from M feature selection classifiers, $f_i = f_{ij}, j = 1, ..., k$
F	Feature space with unique features from M feature lists, $F = f_l^{'}, l = 1, ..., L$
V_l	Voting score for each unique feature where $l = 1, ..., L$

The whole structure of BFSMR is shown in Figure 3.2. After splitting the data into chunks, the Map function is merged with the bootstrapping procedure in Bagging. For each chunk, first it is randomly split into trainset and testset. Given M learning models, M bootstrapped samples are drawn from the trainset with a set ID *sid*. In addition, a testset is drawn ($sid = 0$) without being influenced by any process in the trainset.

The sets with the same *sid* are merged to be used as the inputs of the Reduce function. The Reduce function works in parallel to each group. The original MapReduce normally applies the same model or function to all

groups, but in BFSMR, we match the models with groups based on *sid* and model ID, so that the different model is applied to corresponding groups, which guarantees the possibility of using different feature selection methods to avoid model bias. The outputs of Reducing are M feature lists with K features selected by M models, and the model performance. A feature set F is the union of M feature lists.

Voting strategy is applied to merge the outputs from different models. The voting strategy of Bagging is majority voting with equal probability while we assign feature weights (*w1*) based on the ranking and assign model weights (*w2*) based on their predictive performance. The joint weight is calculated as below. Voting with weights is calculated as Equation 3.4 and the top K features with the highest votes are selected as the final results.

$$w_{jk} = w1_k \times w2_j,\ j = 1, ..., M,\ k = 1, ..., K \tag{3.3}$$

$$v_l = \sum_{j=1}^{M}\sum_{k=1}^{K} v_{jk}w_{jk},\ where\ v_{jk} = \begin{cases} 1, \text{if there is } f_l == f_{jk} \text{ in F} \\ 0, otherwise \end{cases} \tag{3.4}$$

3.2.4 Experimental setup

The data was imported in chunks with a size of 10,000 rows and we got 148 chunks in total with 147 full chunks and the last chunk only including 8857 rows. The split ratio of trainset and testset was 0.8:0.2 and the size of bootstrapped random samples was 10% of the trainset.

The five feature selection methods were applied in parallel and we selected 10 features from each method. We used the nearest-neighbor method to estimate MI in the Filter method (number of neighbors=3) [91]. The features that had the maximum MI with the outcome, were regarded as features with the highest importance. To avoid the problem of long execution time, linear SVM was applied as the estimator for SVM-RFE and the absolute value of

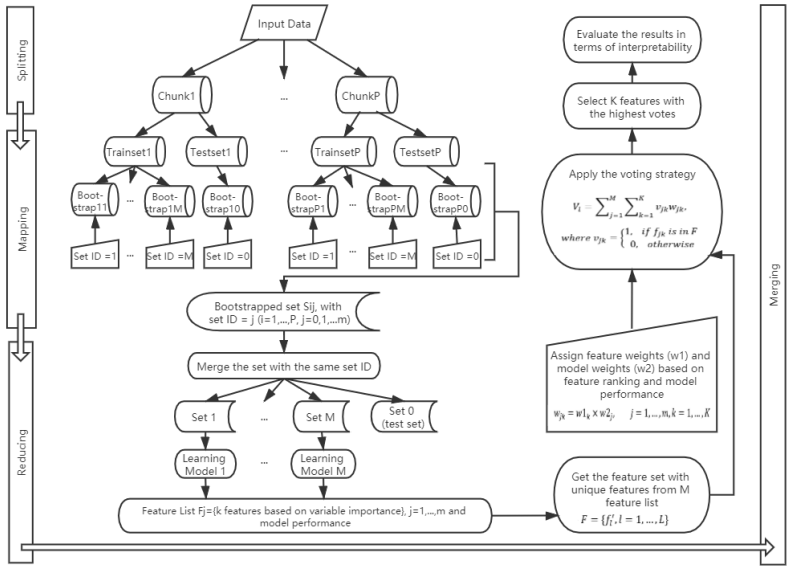


Figure 3.2: The main procedures of BFSMR. There are mainly four steps, i) splitting, ii) mapping which draws bootstrapped samples from each chunk, iii) reducing which merges sets with the same Set ID and applies one feature selection classifier on each set and derive one selected feature list, and iv) merge procedure which combines the selected feature lists into the final output.

the coefficient was the feature importance. The step of the RFE method was set to 1, meaning that one variable was dropped in each iteration and the final 10 variables left in the model were the selected results. To determine the regularization parameter (λ) for Lasso and Ridge regression, the models were iteratively fitted along the regularization path on a grid of parameter, and the parameters that led to the best performance in the cross-validation test were selected, which was 0.002237 for the Lasso regression and 10 for the Ridge regression. As for the Random Forest, we used 50 estimators when training the classifier and used Gini impurity to measure the quality of a split [28]. The most commonly used feature importance is Mean Decrease in Impurity (MDI), however, impurity-based

importance is biased towards numeric features or categorical features with high cardinality [170]. To overcome this limitation, we used permutation importance for feature evaluation [170]. All models were tested on the same testset. As the Filter method selected features without learning algorithms, linear regression was applied using the selected features to get the predictive performance.

The model applies the voting strategy that takes both feature weights and method weights into consideration. The feature ranking from the individual feature selection method was assigned as feature weights (*w1*). The five models were classified into 3 ranks based on model performance (details in Results) and the model weight was set as 1, 0.5 and 0.2 for each rank respectively (*w2*). To examine whether this voting strategy could effectively select the feature set with better interpretability and clinical relevance, it was compared with another two voting strategies and the selected results were compared [182, 173, 98]. These three voting strategies are separate experiments and had no effects on each other.

- Voting1: Voting with equal score for all features
- Voting2: Voting with feature weights
- Voting3: Voting with both feature and method weights

3.2.5 Evaluation measures

It is common to use performance measures to evaluate a learning system, however, in our use case, we would like to mainly evaluate the results in terms of interpretability.

There is no consensus about what interpretability is in machine learning at the current stage [48]. The results of feature selection methods are the potential risk factors of a disease, which should already be self-explanatory

without an explanation method. However, it still makes difference if features have better clinical relevance and are easier to interpret and implement in practice.

To our knowledge, there is no discussion about the dimensions of interpretability of algorithm-identified features. The definitions, properties, and requirements of AI-explanations can show some potential dimensionalities. Properties of explanations which could lead to good interpretability include comprehensibility, certainty, importance, novelty, and representativeness [146], and human-friendly characteristics of explanations are contrastiveness, selectivity, social, inclination to abnormal, truthful, consistent with prior belief, general and probable [110].

After summarizing others' work, we defined the dimensions of interpretability of algorithm-identified risk factors and evaluate the interpretability in Discussion.

3.3 Results

Table 3.1 presents the top 10 features selected from different models. Some variables had negative effects on the outcome indicators, meaning that higher values in the variables indicated lower risk of obesity. These variables are noted by (-) in Table 3.1. There are some common features, such as age, sex, and mother's diet education. However, the different model preference could still be observed. Lasso and Ridge were the specializations of linear regression with different regularization method, thus, they selected similar features, including smoking habits, exercise habits, and diet knowledge. Filter showed a preference for numeric variables (9 out of 10) and they were also variables with fewer missing values. Similarly, Random Forest had an inclination to numeric variables (4 out of 10) and variables with fewer missing values (Sex), but it also selected variables about smoking habits,

exercise habits, and alcohol use. SVM-RFE selected the most different set from the others. Although it also covered diet information and Mothers' exercise habit, the choices of the exact variables were different.

Table 3.1: Top 10 features selected from different models based on variable importance

	Filter (MI)	SVM-RFE	Ridge	Lasso	RandomForest
1	Age	MoDietEducation	Age	Age	SystolicPressure
2	Sleep-Normal (-)	MoRDType-LowSalt	Sex (-)	Sex (-)	MoDiastolicPressure (-)
3	BFType-Maternal (-)	RDType-2000 cal	Tobacco-No (-)	Tobacco-No (-)	MoSystolicPressure (-)
4	DiastolicPressure (-)	AdeDKnowledge	DietEducation	DietEducation	Sex
5	MoSystolicPressure	MoPE-Inadequate (-)	MoTobacco-Yes	MoDietEducation	Birthyear (-)
6	MoNumberCigarettes	DietCompliesAdvice	BFType-Maternal (-)	BFType-Maternal (-)	Tobacco-No (-)
7	Birthheight (-)	MoRDType-Free (-)	PE-Inadequate	Birthyear (-)	MoExerciseAdvice (-)
8	MoBMI	MoPEHour	MoDiabetes-No (-)	MoNumberCigarettes	MoAlcohol-No (-)
9	Birthweight (-)	DiastolicPressure (-)	PE-Adequate(-)	PE-Inadequate	PE-Inadequate
10	MoDiastolicPressure (-)	SystolicPressure	MoDietEducation	DCExecution -No	MoTobacco-Ex

Note: All "Mothers-" in the variables were replaced with "Mo-" for shorter names. Feature abbreviation: RDType - RecommendedDietType, MoRDType – MoRecommendedDitetType, BFType - BreastfeedingType, PE – PhysicalExercise, MoPE – MoPhysicalExercise, MoPEHour – MoPhysicalExerciseHour, AdeDKnowledge – AdequateDietaryKnowledge, DCExecution – DietCorrectExecution.

The model predictive performance was evaluated on the same testset (Table 3.2) based on accuracy and F-score. Ridge, Lasso and Filter had good performance with both measures, SVM-RFE performed at a moderate level as the F-score was not high, and Random Forest had the lowest scores of both measures. Although the performance of Filter was relatively good, it failed to consider variable interactions. Based on the performance and model property, the five models were classified into three ranks with different weights: i) Rank1, Lasso and Ridge (weight=1); ii) Rank2, Filter and SVM-RFE (weight=0.5); iii) Rank3, Random Forest (weight=0.2).

Table 3.2: Comparison of predictive performance among different models.

	Filter (MI)	SVM-RFE	Ridge	Lasso	Random Forest
Accuracy	0.843	0.845	0.844	0.839	0.828
F Score	0.915	0.774	0.915	0.912	0.770

The voting scores for 3 voting strategies are shown in Table 3.3 and visualized in Figure 3.3, which shows the percentage of the score of one feature out of

the whole set. The voting scores can be regarded as the variable importance of this ensemble feature selection framework. Some features gradually gained more importance from Voting1 to Voting3, including age, sex, no smoking, child's diet education, and maternal breastfeeding, which were the top 5 features selected by Voting3. In contrast, mother's diastolic pressure lost its superiority. Inadequate physical exercise had a similar trend although it was still among the top 10 features of Voting3. Mother;s diet education was more stable and took almost the same share of voting scores in all strategies.

The top 10 features selected by BFSMR are age, sex, birth year, breastfeeding type, smoking habit and diet-related knowledge of both children and mothers, exercise, and Mother' s systolic blood pressure. The results indicate that smoking habit, lack of exercise, and unbalanced diet of both mothers and children are the risk factors of childhood obesity. Besides, boys have higher risk than girls and the risk grows along with age. It is also found that maternal breastfeeding can reduce the risk and younger generation (based on birth year) tend to suffer more from obesity probably caused by the changes in the lifestyle, as younger generation tends to have more high-calorie food and spend more time in using smart devices.

3.4 Discussion

3.4.1 Model Interpretability and Clinical Relevance

Clinical relevance and importance

The voting strategy played an important role in selecting a more reasonable feature set with better clinical relevance. When applying individual feature selection methods, there could be bias in the results caused by model limitation. For example, as shown in Table 1, Filter and Random Forest

Table 3.3: Comparison of the selected variables with high scores calculated from different voting strategies

Voting1		Voting2		Voting3	
PE-Inadequate	4	Age	30	Age	25
Age	3	Sex	25	Sex	19.4
BFType-Maternal	3	Tobacco-No	21	Tobacco-No	17
MoDietEducation	3	BFType-Maternal	18	BFType-Maternal	14
Sex	3	MoDietEducation	17	DietEducation	14
Tobacco-No	3	DietEducation	14	MoDietEducation	12
Birthyear	2	MoSystolicPressure	14	PE-Inadequate	8.4
DiastolicPressure	2	SystolicPressure	11	MoTobacco-Yes	6
DietEducation	2	Birthyear	10	MoNumberCigarett	5.5
MoDiastolicPressure	2	MoDiastolicPressure	10	Birthyear	5.2
MoNumberCigarettes	2	PE-Inadequate	10	MoSystolPressure	4.6
MoSystolicPressure	2	DiastolicPressure	9	DiastolicPressure	4.5
SystolicPressure	2	MoRDType-LowSalt	9	MoRDType-LowSalt	4.5
AdeDKnowledge	1	Sleep-Normal	9	Sleep-Normal	4.5
Birthheight	1	MoNumberCigarettes	8	RDType-2.000 cal	4
Birthweight	1	RDType-2.000 cal	8	AdeDKnowledge	3.5
DietCompliesAdvice	1	AdeDKnowledge	7	MoDiabetes-No	3
DCExecution-No	1	MoPE-Inadequate	6	MoPE-Inadequate	3
MoAlcohol-No	1	MoTobacco-Yes	6	DietComplieAdvice	2.5
MoBMI	1	DietCompliesAdvice	5	SystolicPressure	2.5
MoDiabetes-No	1	Birthheight	4	MoDiastoPressure	2.3
MoExerciseAdvice	1	MoExerciseAdvice	4	Birthheight	2
MoPE-Inadequate	1	MoRDType-Free	4	MoRDTyp-Free	2
MoPEHours	1	MoAlcohol-No	3	MoBMI	1.5
MoRDType-LowSalt	1	MoBMI	3	MoPEHours	1.5
MoRDType-Free	1	MoDiabetes-No	3	Birthweight	1
MoTobacco-Yes	1	MoPEHours	3	DCExecution-No	1
MoTobacco-Ex	1	Birthweight	2	MoExerciseAdvic	0.8
RDType-2.000 cal	1	DCExecution-No	1	MoAlcohol-No	0.6
Sleep-Normal	1	MoTobacco-Ex	1	MoTobacco-Ex	0.2

Note: All "Mothers-" in the variables were replaced with "Mo-" for shorter names. Feature abbreviation: RDType - RecommendedDietType, MoRDType – MoRecommendedDitetType, BFType - BreastfeedingType, PE – PhysicalExercise, MoPE – MoPhysicalExercise, MoPEHour – MoPhysicalExerciseHour, AdeDKnowledge – AdequateDietaryKnowledge, DCExecution – DietCorrectExecution.

had an inclination to numeric variables and variables with fewer missing values. In Experimental setup, we explained that we used permutation importance for feature evaluation when applying Random Forest, as it was known that the commonly used feature importance, MDI, would be

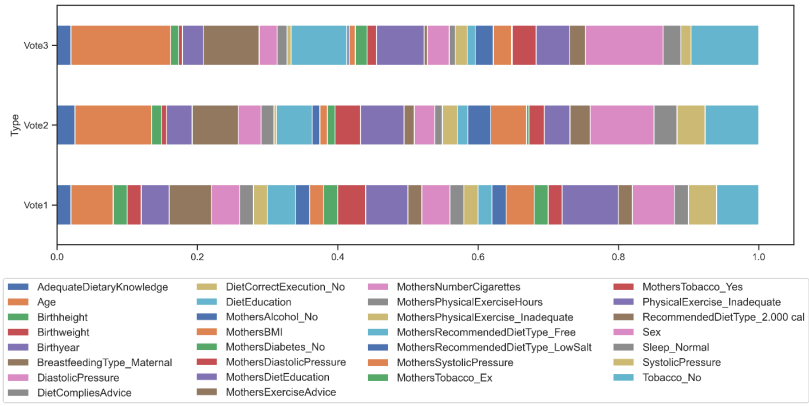


Figure 3.3: Visualization of the comparison of selected variables with high scores calculated from 3 voting strategies. The plot was drawn using the percentage of the score of one feature out of the whole set and larger percentage means higher variable importance. Age, Sex, Tobacco-No, DietEducation, and BreaskfeedingType-Maternal gradually gained more importance during the changes from Voting1 to Voting3 while the importance of MothersDiastolicPressure and PhysicalExercise-Inadequate dropped. MothersDietEducation was more stable and took almost the same share of voting scores in all strategies.

biased towards numeric features or categorical features with high cardinality. However, the results turned out that there was still bias in the results even if we used permutation importance. These results proved that there would be risks that categorical variables with high missing rates were disregarded if some feature selection models were applied individually. Our method, BFSMR, provided combined results from a set of feature selection methods, which could reduce or eliminate the bias affected by missing values.

Bagging uses the majority voting with equal probability (Voting1). But the close scores of Voting1 made it difficult to distinguish the most important ones. For example, 13 features had scores higher than 2, making it impossible to select only the top 10 features as the final output. Voting2 added feature weights based on the ranking from each model, nevertheless, variables with

high rankings from poorly-fitted models could still affect the results. One significant difference between Voting2 and Voting3 was the notable decline in the rankings of the numeric variables of less relevance, e.g. mother's systolic pressure, child's systolic pressure. The relevant numeric variables were not negatively affected, on the contrary, the average number of mother' s cigarette consumption climbed from the 15th in Voting2 to the 9th in Voting3.

Easiness to implement in practice

BFSMR could select variables that were easier to implement in practice for the follow-up interventions or policies. For example, although three models (SVM-RFE, Ridge, Lasso) covered diet-related information, different variables were selected, e.g. diets with less than 2000 cal, whether follow the diet advice, whether correctly execute diet advice. These variables were concrete about one specific aspect but it would be difficult to use them to make corresponding interventions. The features selected by BFSMR were child' s diet education and mother' s diet education , which could be easily put in practice in real life, for example by arranging lectures at school or suggesting pregnant women taking specifically developed courses.

General and probable

BFSMR is a general framework to deal with large-scale data and combine results from multiple models. It is flexible because the individual feature selection methods applied in this framework are not strictly defined but can be replaced with other methods, and more methods of interest can be added into the framework. Therefore, it is probable to apply BFSMR in most use cases.

Representativeness

In the study, 5 models were chosen as representatives of filter, wrapper, and embedded methods, meaning that a wide range of model types were covered. Apart from the methods, the algorithm-identified risk factors are also representative risk factors that can well summarize one aspect, for example, child' s diet education is a good representative for diet-related suggestions.

Comprehensibility

Although the results of individual model were not reported as the final output, the process of applying multiple models as components of ensemble feature selection was a comprehensive exploration of important features. Some variables only appeared once in one model, which might be neglected if only one model was applied. For instance, Sleep-Normal had a negative effect in the Filter method, which was in agreement with clinical knowledge [83, 156]; mother's BMI and mother' s diabetes disease history were reasonable and could indicate genetic influence; mother' s diet type (Low Salt) and mother' s diet type (Free) suggested the ideal diet styles. These variables were not included in the current output, but it would be very likely these variables were selected if we chose a higher number of selected features, for instance, Sleep-Normal and mother' s diet type (Low Salt) ranked 12th and mother' s diabetes disease history was at the 17th.

Consistency with prior beliefs

There are numerous studies on the risk factors for childhood obesity, which can be classified into two main types, the genetic factors and environmental factors [92]. This paper focused more on environmental factors, as it would be easier to make corresponding policies for environmental factors. This

type of factors include lifestyle factors such as eating behaviors [52], sleeping pattern [83, 156], parental feeding styles [52]. The environmental factors may also include some other factors such as environmental chemicals or microbiota, but the database used in this paper does not include such information. A study in 2001 suggest that the main risk factors for obesity in children include dietary intake, physical activity, family characteristics, parents' lifestyles and environmental factors such as school policies and demographics [41]. Another study in 2013 identified parent BMI, child sleep duration and parental restrictive feeding as the main risk factors [46]. A more detailed literature review of previous studies on risk factors of childhood obesity can be found in Section 3.6.2.

Based on the previous studies about the risk factors of childhood obesity (Section 3.6.2), the main causes include lifestyle factors such as eating behaviors, physical activity, sleep, age, gender, parents' lifestyles, and smoking habit, which were all covered in our results. Some studies pointed out the relation with genetic factors and psychological factors, but such information was lacking in our data. In general, our results are consistent with prior beliefs.

3.4.2 Contributions and limitations

One contribution of the paper is the distributed implementation for ensemble feature selection. Although some studies applied distributed implementation for individual methods [142], few papers tried it for ensemble feature selection. A homogeneous-distributed ensemble was proposed [155, 25, 113], which split the trainset into subsets and applied the same method. However, using the same method on the divided subsets cannot avoid model bias. Thus, it is necessary to apply distributed implementation for ensemble feature selection when multiple feature selections are included. Moreover, MapReduce avoided full load, so that the real-world problem of platform

limitation was solved.

Another contribution is the voting strategy. In general, there are three ways of combining the outputs of ensemble feature selection: combination of label predictions, features subsets, and features rankings [24]. The combination of label predictions are most widely used and the models are trained to achieve the best performance. However, this method cannot work well when the goal is to select a subset of features with clinical relevance. To solve the problem, some studies combine the results by having the intersection or union of the features from different selectors [101], and some studies incorporate feature rankings (Voting2) [171, 6]. Our voting strategy (Voting3) has both feature rankings and model performance, making it possible to select a more reasonable set. A detailed literature review is in Section 3.6.2.

Furthermore, no discussion was identified on the interpretability of algorithm-identified risk factors. To better evaluate interpretability, the dimensions of interpretability were defined and summarized in this paper.

Finally, previous studies showed that there were regional differences in childhood obesity trends [131], but there were no studies in more details on a regional level. Basque government' s goal within this study was to find out the risk factors of their own region and make corresponding policies.

One main characteristic of BFSMR is its flexibility. This paper used 5 feature selection methods, namely filter methods based on MI, SVM-RFE, Lasso Regression, Ridge Regression, and Random Forest. However, it is not necessary to apply these methods in all use cases and the number of models can be more or less than 5. The type of feature selection methods can be adjusted according to the specific research goal of the study. Similarly, the model weights can also be adjusted based on the model performance. This paper classified the models into 3 ranks, because it could be observed that the models were on high, moderate, and low levels in terms of performance evaluation. In a different use case, it could happen to evaluate 10 models,

and to adopt a classification of 4 or 5 ranks or 4 models into 2 ranks (the technological approach is flexible). In fact, this ensemble feature selection method is similar to the process of formulating standards and industry guidelines by an expert committee. Each individual method acts as one expert and the model weight can be regarded as the priority of one expert' s opinion because of the expert' s past experience. Therefore, the model weight needs to be discussed case by case based on the real situation, so that we don' t give a strict requirement of the number of ranks or the value of weights. The core values of BFSMR is the distributed implementation for ensemble feature selection method and the voting strategy to merge the final results.

Most clinical studies of childhood obesity focused on smaller age groups to give precise conclusions [14, 119]. Our participants had a large age span, aiming for better policy making. We defined the outcome indicator on a population level, targeting on policy for all children, instead of a specific age group. It is acceptable, as the policy is normally to adjust general behavioral patterns that can have long-term effects, and our results could well serve this purpose. The current results are for Basque Country, subgroup analyses will be done for province-level and town-level data and it will eventually go to small villages.

One limitation of this study is the predictive performance of the five models was not very satisfactory due to the sparse structure, missing values, and the small number of selected features. The model performance can be further improved if more features are kept, or by using methods that are specifically for sparse data. Another limitation is causality. The current study, same as other classical feature selection methods [66], selected features based on correlation, and the models were diagnostic models instead of prognostic prediction models [153]. However, it could be solved if the individual method within the framework was causality-based models or prognostic models. In summary, BFSMR acts as a general strategy to provide the framework of a

meta-algorithm. The feature selection models can be replaced if particular data issues need to be solved. In terms of generalization, one limitation of this study is that all the participants were the children from Basque region. Previous studies showed that there were regional differences in childhood obesity trends,[41] thus, the generalization of the identified risk factors to other populations should be cautious.

3.5 Conclusion

We presented a new ensemble feature selection framework that combines MapReduce and Bagging to make it possible to deal with large-scale data and applies 5 feature selection models to avoid model bias. A collection of risk factors of childhood obesity with better interpretability and clinical relevance were identified, which solved the practical research question raised by the department of health, to contrast with their experience and knowledge and provide supports for the decision-making of future interventions and policies. The proposed framework can also be applied to select risk factors of other chronic conditions in the future.

3.6 Supplementary Materials

3.6.1 Data pre-processing

Data cleansing and harmonization

The Osakidetza dataset is composed of data extracted from primary care, specialty care and hospital databases, including information about the patient, diagnosis, medical forms, medical appointments, prescription information, and information about the child' s mother. There are 9

datasets in Osakidetza. The datasets extracted from Osakidetza needed cleansing since they had inconsistencies, empty columns, and incorrect formatting values. In addition, data about the forms were originally stored in five different datasets: three about children' s forms taken from primary care, specialty care and nurse collection records respectively and two about mothers' forms taken from primary care and specialty care.

First, data harmonization was conducted as different form sources contain variables that refer to the same concept but expressed in different ways and/or in different units (e.g. the "weight" variable was expressed in grams on primary care forms and the "weight" variable was expressed in kilograms on specialty care forms). Second, the five forms datasets were merged into two datasets (i.e. children' s forms and mothers' forms) based on childrenID and motherID. Third, the mother' s forms data was aggregated by year by using the mean values for the numeric variables and the mode values for the categorical variables and dummy variables with values of 0 and 1. The variables of sex, birthdate, and the child' s mother' s ID in children' s information table were added into children' s form data and this merged table was further joined with mother' s aggregated forms data based on motherID and registration year in both tables. In this case, the mother' s forms data becomes additional variables for the children to provide maternal and hereditary information. Finally, the empty columns were dropped and rows with empty BMI were not useful for our study. In the end, only records of the children under the age of 18 with children' s BMI in the range between 10 and 60 were kept.

The final merged data has 1,478,857 records from 426,813 children.

Data pre-processing

The outcome was defined based on BMI. In fact, other variables such as waist are also used as indicators in obesity-related studies [78], which will

show strong correlation with our outcome if these variables are included in the predictors. In this case, 5 variables were dropped, i.e. BMI, height, weight, waist, and size.

There are three types of variables per record: 28 numeric variables, 8 dummy variables, and 33 categorical variables. Almost all variables have missing values. We replaced missing values with 0 for the numeric variables and with "Unclear" for categorical variables. With the refilled "Unclear", the n-level categorical variables were converted into n+1 dummy variables by using One-Hot encoding [5]. To overcome the limitation of One-Hot encoding, the dummy of "Unclear" was dropped to avoid the multicollinearity [64]. It was then used as the default option for each variable. Eventually, 83 dummy variables replaced 33 categorical variables, with each dummy representing one level of the categorical variables, denoted as "VarName-LevelName". The exception was the "Yes" dummy of the two-level categorical variables, denoted as "VarName". Both "Yes" and "No" dummy were kept for the two-level categorical variables, because there were a considerable number of "Unclear" observations, meaning that "Yes" and "No" didn't represent the same information and it would not cause multimorbidity when keeping both of them. It was tested to process the eight dummy variables in the same way as categorical variables. However, it would lead to severe multicollinearity in the data because the dummy variables had limited negative observations. In this case, we only kept the dummy representing the positive answers. All the data was standardized before model application. Finally, we had 119 variables in total and the detailed variable names can be found in Table 6.3.

3.6.2 Related Work

Risk factors of childhood obesity

There are numerous studies on the risk factors for childhood obesity. In general, obesity is often considered to be the result of an imbalance between calories taken in and burned out. However, there is increasing evidence that other factors such as genetic background play a key role in determining the risk of obesity [21]. According to the review in 2017 [92], childhood obesity is the result of an interaction between different factors such as the environment, genetics and a child' s surroundings. Environmental factors include lifestyle factors such as eating behaviors, which is highly related to parents' feeding styles, physical activity, stress and depression[52]. Other major environmental factors include perinatal factors [40, 96], birth size [199], catch-up growth [177], environmental chemicals [193], microbiota [85, 29], and adverse life experiences [62]. There are some studies that have demonstrated an association between the sleep time duration and obesity [83, 156].

A study in 2001 suggest that the main risk factors for obesity in children include dietary intake, physical activity and sedentary behaviour, moderated by factors such age and gender. In addition, family characteristics, parents' lifestyles and environmental factors (e.g. school policies and demographics) have a major impact in children' s lifestyle and, therefore, their risk of obesity [41]. Another study in 2013 used multiple regression analyses to identify childhood obesity risk factors from data collected in a longitudinal study of preschool children and they conclude that the three early-life risk factors are parent BMI, child sleep duration and parental restrictive feeding [46]. Hammond' s research predicted childhood obesity using electronic health records and publicly available data by means of a variety of machine learning algorithms [69].

It is expected that the selected features of our framework could be in corresponding to previous studies, proving the reliability of the results and the possibility of using the framework to select risk factors in the future study.

Combination of the results for ensemble feature selection

In general, there are three ways of combining the outputs of ensemble feature selection: combination of label predictions, combination of subsets of features, and combination of rankings of features [24]. The combination of label predictions are most widely used in studies, which use techniques such as majority voting to select the most likely class labels from the class labels predicted by all classifiers, or use decision rules to decide the final class label if the classifiers output a degree of certainty [132]. The final outputs are the predicted class labels, and performance measures, e.g. accuracy or precision, are often used to evaluate the results.

However, in some cases, it is expected to have selected features as the output and the goal is to select a subset of features that are diverse, stable, and interpretable. If classification is still necessary, it can be applied as an additional procedure. The most typical way of combining the subsets of features is to use the intersection or union of the features selected by different selectors [101]. Some studies incorporated decision rules when making unions, e.g., they only included the new subsets into the union when the new subsets could improve the classification accuracy [25], or data complexity [113].

The last way of combining the results is to combine the rankings of the features, as some feature selection methods are capable to output an ordered feature list, instead of a set of features. The rankings of the features can be aggregated by using some simple statistics, such as minimum or median. Some studies improved the aggregation method by using order statistics

[171, 6]. They compared the actual rankings with the expected behavior of uncorrelated rankings, and then re-ranked the features and calculated significance scores. However, this method requires simulations to define significance thresholds and does not support partial rankings, meaning that the rankings must contain all the features.

The last two methods of combining results share similarity as both of them return the selected features instead of the predicted labels. Therefore, it is worth to try a joint method to take the advantages of these two methods.

Table 3.4: A full list of features used in the experiment.

Numeric Variables (28)	Birthyear, Age, SystolicPressure, DiastolicPressure, Glucose, Triglycerides, UricAcid, Gpt, Got, FruitsPerDay, FruitVegetableConsumption, NumberCigarettes, AlcoholUnitWeek, Birthheight, Birthweight, GestationalAge, CardiovascularRisk, FraminghamCardiovascularRisk, RegicoCardiovascularRisk, MoSystolicPressure, MoDiastolicPressure, MoBMI, MoAlcoholUnitWeek, MoNumberCigarette, MoCardiovascularRisk, MoFraminghamCardiovascularRisk, MoRegicoCardiovascularRisk, MoPhysicalExerciseHours
Dummy Variables (8)	Sex, DietIntentChange, DietCompliesAdvice, MoRecommendedDiet, MoBirth, MoGestationalDiabetes, MoDietIntentChange, MoUnknownVariable
Categorical Variables (33/83)	**YES/NO(19/38)**: Alcohol, Diabetes, ExerciseAdvice, BreakfastDairy, BreakfastFruit, DietCorrectExecution, AdequateDietaryKnowledge, MoAlcohol, MoExerciseAdvice, MoFitnessAdvice, MoDiabetes, MoDietCorrectExecution, MoAdequateDietaryKnowledge, MoBirthPreparation, MoBreastfeedingEducation, MoMaternalBreastfeedingInformation, BreastfeedingAbandonment, DietEducation, MoDietEducation; **Normal/Abnormal(1/2)**: Sleep; **Adequate/Inadequate(4/8)**: Diet, MoDiet, PhysicalExercise, MoPhysicalExercise; **Multiple categories (9/45)**: Tobacco(6), MoTobacco(6), MoSmoker(2), RecommendedDietType(6), MoRecommendedDietType(8), TypeBreastfeeding(4), MoTypeBirth(6), MoPlaceBirth(4), MoPromotionBreastfeeding(3)

Chapter 4

Predictive Models for Chronic Disease

This chapter is submitted to Frontiers in Medicine as:

Shi X, Qu T, Van Pottelbergh G, van den Akker M, De Moor B. A prognostic model for End-Stage Renal Disease: an analysis of Belgium primary care data.

ABSTRACT

Background: Prognostic models can help to identify patients at risk for End-Stage-Renal Disease (ESRD) at an earlier stage to provide preventive medical interventions. Previous studies mostly applied Cox proportional hazards model. The aim of this study is to present a resampling method, which can deal with imbalanced data structure for the prognostic model and help improve predictive performance.

Methods: The Electronic Health Records of Chronic Kidney Disease (CKD)

patients older than 50 during 2005-2015 collected from primary care in Belgium were used (n=11,645). Both Cox model and logistic regression were applied as reference model. Then the resampling method, SMOTE-ENN, was applied as a preprocessing procedure followed by the logistic regression. The performance was evaluated by accuracy, AUC, confusion matrix, and F_3 score.

Results: The C statistics for Cox model was 0.807 while the AUC for logistic regression was 0.700, both on a comparable level to previous studies. With the model trained on the resampled set, 86.3% of ESRD patients were correctly identified. The F_3 score was 0.245, much higher than 0.043 for logistic regression and 0.022 for Cox model.

Conclusions: Our study pointed out the imbalanced data structure and its effects on prediction accuracy, which were not thoroughly discussed in previous studies. We were able to identify patients with high risk for ESRD better from a clinical perspective by using the resampling method. The technique can be widely used in other clinical topics when imbalanced data structure should be considered.

Key words: Logistic regression, resampling method, predictive performance, chronic diseases

4.1 Introduction

Chronic kidney disease (CKD) is a common chronic disease with high prevalence ranging from 5% to 25%, according to the survey studies in European countries [166]. Of all patients with CKD, only a small proportion will progress to End-Stage Renal Disease (ESRD) when dialysis or kidney transplants are necessary [86, 181]. ESRD is a real burden for the patient with the high mortality rate and the high cost of renal replacement therapy [111]. The evolution from the early stage of CKD to ESRD takes several

years in most case [137]. The guidelines of CKD based on sufficient evidence show that good follow-up treatments of CKD can slow down the progression towards ESRD [168]. Therefore, an accurate predictive method that can identify the CKD patients with the high risk of ESRD at an earlier stage is essential to slow down the progression. A systematic review on prognostic models for CKD shows that previous studies mostly applied Cox proportional hazards model (87%) or logistic regression (3%) [190], and this is also true for papers on prognostic models for ESRD [143, 95].

Despite the low prevalence of many chronic diseases, for instance in our use case, only a small portion of CKD patients will progress to ESRD, the impact of this low prevalence on the validity of prognostic models have not been thoroughly discussed in previous studies. The low prevalence, e.g. 15% developing ESRD vs 85% not developing ESRD, raises the necessity of discussion about imbalanced data structure. Imbalanced data is a dataset with a skewed distribution. Classes with a large proportion (in our case people with CKD not developing ESRD) are called majority, while classes with a small proportion (those developing ESRD) are called minority. The degree of imbalance is regarded as moderate if the proportion of minority class is within the range 1%-20% [4]. The prevalence of most chronic diseases are within this range, therefore, the imbalanced data structure is a data issue that requires attention for clinical studies.

The biggest problem caused by imbalanced data structure is called "accuracy paradox", which describes a situation where accuracy is unable to reflect the real performance of a model [196, 187]. A model will tend to predict every case as category A if the incidence of category A is dominant, e.g. 99%, leading to a perfect accuracy of 99%. In the context of clinical studies, this means the model predicts every participant as no risk of developing a disease in the future to achieve good performance. Not only accuracy, but also some other commonly used evaluation measures, such as precision, recall, AUC, and C statistics for survival models, are not

adequate performance metrics for imbalanced datasets [16]. Therefore, it is necessary to provide a feasible solution which can deal with imbalanced data structure, improve the model' s ability to identify minority cases, and evaluate the performance correctly.

This study aims to introduce the use of resampling methods as data preprocessing to improve the predictive accuracy of high-risk patients for an evolution towards ESRD and suggest an evaluation measure with better clinical relevance.

4.2 Materials and Methods

4.2.1 Patients

The study used Intego, a longitudinal database collecting electronic medical records of patients in general practice in Flanders Belgium since 1994. Data is yearly collected from the daily consultation, including every patient who had a contact with their GP in that year. About 300,000 individual patients are recorded in the database, corresponding to over 2.3% of the Flemish population and this Intego population is representative for the general population in Flanders, Belgium [184]. Intego contains all the coded data registered in general practices, including clinical parameters, laboratory tests, disease diagnosis and prescriptions. The diagnosis are coded based on the International Classification of Primary Care -2 (ICPC-2) and prescriptions are based on the Anatomical Therapeutic Chemical (ATC) classification system. Intego uses an opt-out methodology and is approved by the local ethical committee of the KU Leuven and in line with Belgian and European privacy regulations.

The target population for the study is CKD patients older than 50 years old with records from 2005 to 2015. A total of 11,645 patients were

involved in our study, including 11,424 CKD patients and 221 patients who developed ESRD. Among all the patients, the average interval between the last observation and its former records was 1.358 years, meaning that our study could be broadly regarded as 1.5-year predictive model.

4.2.2 ESRD and Risk Factors

To define CKD patients and ESRD patients, we used the eGFR as recommended [167].

$$eGFR = \begin{cases} 175 * creatinine^{-1.154} * age^{-0.203} & if male \\ 0.742 * 175 * creatinine^{-1.154} * age^{-0.203} & if female \end{cases} \tag{4.1}$$

CKD patients were defined as patients with eGFR lower than $60mL/min/1.73m^2$ for three or more months. Patients with CKD were labeled as having ESRD if they had an eGFR once lower than $15mL/min/1.73m^2$ and the mean of the last two measurements was lower than $30mL/min/1.73m^2$ [137]. Each time point would be recorded in a new variable *ESRD Time*, which tracked the progression of CKD. In the follow-up prediction analysis, the end point could be patients having ESRD for the first time or the end of the observation.

The risk factors were mainly selected referring to previous studies[137, 172, 202, 35, 186], including Age, Gender, Hemoglobin, Uric Acid, Hypercholesterolaemia, Type 2 Diabetes Mellitus, Obesity, Hypertension and Malignancy. Apart from these variables which were widely used in ESRD studies, we added two types of medication, antihypertensive medication and cholesterol-lowering medication. Previous studies show the decline in blood pressure is a strong risk factor in the decline of kidney function [186, 45], and the use of antihypertensive medications may have accelerated kidney damage [90]. Besides, cholesterol-lowering medication decreases the overall

risk of CVD and mortality of CKD patients [140]. Hence, we included these medications in our study. Some risk factors considered in other studies were not included due to lack of sufficient information in the database, for example tobacco use. The detailed description of risk factors can be found in Table 4.5 in Section 4.6.2.

4.2.3 Data Preprocessing - Resampling Methods

In this study, the resampling method was applied as a preprocessing step to solve the data issue of imbalanced structure due to the prevalence. SMOTE-ENN was used in this study after testing a collection of resampling method (The tested methods are listed in Section 4.6.1) [17], as it was the only one which could deal with the imbalanced structure without losing the original pattern and could yield the best performance when applied together with the predictive model.

SMOTE-ENN is a combination of oversampling and undersampling methods, meaning that it increases the number of minority class first and then reduces the number of majority class. The algorithm functions with the target of adding and removing observations without changing the distributions of the majority and minority class. One characteristic of SMOTE-ENN is that the resampled data does not have the same numbers in different classes. Instead, the resampled majority class is still larger than the minorities, but the gap in size between different classes is much closer. For example, a variable may have 100 values as *CategoryA* and 900 values as *CategoryB*. After applying SMOTE-ENN, the resampled data may have 480 values as *CategoryA* and 670 values as *CategoryB*. Other resampling methods tend to have equal numbers in different categories, i.e., both categories have 900 values after applying oversampling method, or both have 100 values after applying undersampling method. This preprocessing procedure was done in Python 3.6, imblearn 0.5.0.

4.2.4 Statistical Analysis

We first compared the baseline characteristics of CKD and ESRD patients. The mean and standard deviation of continuous variables and the percentage of binary variables were calculated. The chi-square test was conducted to check whether these variables were statistically different for CKD and ESRD patients.

The whole data was randomly split into a train set and a test set at the ratio of 0.75:0.25 based on the patient ID. All the records in the train set were used to learn the characteristics of ESRD patients, while in the test set, the records in the last observation year for non-ESRD patients and the records with ESRD diagnosis for ESRD patients were removed to create a scenario where the future status was kept unknown.

Logistic regression and Cox proportional hazards model were applied as reference models. The Odds Ratio (OR) for logistic regression and Hazard Ratio (HR) for Cox model were calculated respectively, and 95% confidence interval (CI) were given to show the significance level of the estimated parameters.

For Cox model, time-to-event outcomes were censored for loss to follow-up or end of study. For the logistic regression, the outcome was an indicator of whether to progress to ESRD during the observation period and it was defined based on the patient’s whole history in the database. The patients could have multiple records at different periods and all the records were kept without selection. Records from patients with no ESRD occurrence would be labeled as “0”. For a given patient with at least one ESRD occurrence, “1” was assigned to all records, although some records were at the early stage when he had not developed ESRD yet. We defined the outcome in this way so that the outcome was not to represent the temporal status of the patients, instead, it classified the participants into two groups, patients who would not have ESRD and patients who would. The calculation of

logistic regression and Cox model was done in R 3.5.1.

In the next step, SMOTE-ENN was applied on the train set for the outcome to generate a resampled data including less observations in majority class and more in minority class, aiming to improve the prediction performance in case of unbalanced data structure. Then logistic regression was applied again on the resampled data and the performance was evaluated on the original test set and compared with logistic regression without resampling methods and Cox model.

Accuracy and AUC (Area under the ROC Curve) were the main measures to assess the prediction performance in previous studies and C statistic was used for survival models. Confusion matrix is a table to describe the performance of the classification problem by dividing the results into four groups, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). As the aim of the study was to detect positive cases among all participants, it was more important to avoid labeling patients of high risk as negative cases than monitoring a low-risk patient for a period. In this case, we used another measure, F_β, which was a combination of precision and recall with different weights. The F_β-score is a series of scores constructed following a formula, where the parameter β represents the level of penalty of False Negative. The parameter was set as 3 in our study.

$$F_\beta = \frac{((1+\beta^2) * Precision * Recall)}{((\beta^2 * Precision + Recall))} \tag{4.2}$$

The estimated outcome of logistic regression was the probability of developing ESRD while that of Cox model was survival probability. The default thresholds to define positive and negative cases for both Cox model and logistic regression were 0.5. To examine the effects of threshold choice on the predictive performance, the confusion matrices and other evaluation measures of logistic regression (threshold=0.7) and Cox model (threshold=0.05 and 0.01) were calculated.

4.3 Results

4.3.1 Study Participants

A total of 11,645 patients were involved in our study, including 11,424 CKD patients and 221 ESRD patients. Table 4.1 summarizes the detailed baseline characteristics of the CKD and ESRD patients. The prevalence of ESRD among all study participants was 1.90%. Based on the statistical test, *Agents acting on the renin-angiotensin system, Beta blocking agents, Age*, and *Type 2 Diabetes Mellitus* were not statistically significantly different for CKD and ESRD patients.

Table 4.1: Baseline characteristics of CKD patients over 50 years old in Intego (2005-2015) by future progression or no progression to ESRD

	CKD	ESRD	p-value
No. of Observation (No. of patients)	58529 (11424)	2257 (221)	
Hemoglobin (mean ± SD)	13.2 ± 1.68	11.9 ± 1.60	<0.001
Uric Acid (mean ± SD)	6.5 ± 1.81	7.2 ± 2.18	<0.001
Age (mean ± SD)	75.3 ± 9.66	75.5 ± 9.59	0.46
Type 2 Diabetes Mellitus (%)	14.3	13	0.09
Hypercholesterolaemia (%)	10.4	6	<0.001
Use of Antihypertensives drugs (%)	3.9	14.9	<0.001
Use of Diuretics (%)	30.1	35.7	<0.001
Use of Beta blocking agents (%)	41.6	41.4	0.88
Use of Calcium channel blockers (%)	17.9	31.7	<0.001
Use of Agents acting on the renin-angiotensin system (%)	41.8	41.9	0.98
Use of Lipid Modifying Agents (%)	42.4	38	<0.001
Obesity (%)	1.4	0.6	0.001
Hypertension (%)	15.2	11.9	<0.001
Malignancy (%)	17.3	24	<0.001
Male gender (%)	60	53.2	<0.001

4.3.2 Resampling Methods

The train set and test set were divided based on the patient ID at the ratio of 0.75:0.25, so the number of patients in train set was 75% of the total participants, but the number of observations was not strictly 75% of the

total observation (44,006 CKD records and 1,631 ESRD records). After the resampling procedure, the number of CKD observations slightly decreased, while that of ESRD records grew to a large number as high as more than 26.5 times the original number (39,110 CKD records and 43,232 ESRD records).

4.3.3 Estimates of three models

The estimated parameters of three models are shown in Table 4.2. All the estimates of the coefficients were statistically significant and *Antihypertensives, Calcium channel blockers, Malignancy,* and *Uric Acid* had odds ratio higher than 1, meaning that they had positive effects on the probability of developing ESRD, while odds ratios lower than 1 indicated negative relationship.

Table 4.2: Parameter estimates of Logistic Regression and Cox Model to predict the risk of future progression to ESRD

	SMOTE-ENN+Logistic Regression		Logistic Regression		Cox Model	
	OR	95% CI	OR	95% CI	HR	95% CI
Hemoglobin	0.501	0.495-0.507	0.587	0.561-0.616	0.64	0.612-0.670
Uric Acid	1.189	1.178-1.199	1.088	1.040-1.137	1.055	1.010-1.102
Age	0.978	0.976-0.979	0.976	0.967-0.986	0.967	0.958-0.976
Type 2 Diabetes Mellitus	0.777	0.740-0.817	0.849	0.637-1.114	0.711	0.541-0.933
Hypercholesterolaemia	0.624	0.585-0.666	0.661	0.431-0.971	0.59	0.396-0.877
Male gender	0.55	0.531-0.569	1.341	1.110-1.619	1.211	1.007-1.456
Diuretics	3.095	2.884-3.323	2.684	1.963-3.618	2.697	2.005-3.628
Beta blocking agents	0.862	0.830-0.897	1.121	0.907-1.381	1.268	1.034-1.555
Calcium channel blockers	0.962	0.927-0.998	0.89	0.719-1.101	0.847	0.687-1.045
Agents acting on the renin-angiotensin system	1.844	1.768-1.924	1.99	1.581-2.493	1.935	1.551-2.415
Diuretics	0.885	0.853-0.918	0.429	0.341-0.538	0.443	0.355-0.553
Lipid Modifying Agents	0.761	0.733-0.791	0.865	0.693-1.077	0.722	0.581-0.897
Obesity	0.142	0.103-0.192	0.215	0.012-0.969	0.15	0.021-1.070
Hypertension	0.668	0.636-0.702	0.749	0.548-1.003	0.598	0.444-0.805
Malignancy	1.385	1.330-1.442	1.015	0.803-1.272	0.875	0.699-1.096

Note: OR = Odds Ratio, HR = Hazard Ratio, CI = Confidence Interval

The results of logistic regression of resampled data was different from results of logistic regression and Cox model of original train set. There were two more variables, i.e. *Gender (Male)* and *Diuretics*, indicating positive effects in the latter two models. It was also possible that the values of odds ratios or hazards ratios differed even though they all meant positive or negative

relations. For example, the value of estimates of *Hemoglobin* gradually increased from 0.501 for SMOTE-ENN+Logistic Regression, to 0.587 for Logistic Regression, and to 0.640 for Cox Model. The similar increasing or decreasing trend among these three models could be observed in *Uric Acid, Diuretics, Beta blocking agents*, and *Malignancy*. Besides, the estimate of odds ratio in SMOTE-ENN+Logistic Regression could have distinct values when the other two models had relatively the same values. For instance, *Gender (Male), Antihypertensives, Diuretics* and *Agents acting on the renin-angiotensin system*. Although the values of *Agents acting on the renin-angiotensin system* in three models were all lower than 1, the estimated value of SMOTE-ENN+Logistic Regression was almost two fold of the other two values. In some cases, Logistic Regression had the highest values among the three models, such as *Diabetes, Hypercholesterolaemia, Calcium channel blockers, Lipid modifying agents, Obesity*, and *Hypertension*, considering most of these variables were not significant in Logistic Regression, we could interpret this type of difference as limitation of Logistic Regression to correctly estimate the results.

4.3.4 Performance evaluation

The performance of all three models was evaluated on the original test set. There were 51 ESRD patients and 2858 CKD patients in the test set. We made a prediction for the patients about their possibility to develop ESRD in the future state, therefore, there was one result for one patient as an overall prediction.

Table 4.3 shows the confusion matrices of three models and the threshold to decide whether the patients would develop ESRD or not were 0.5. Although the accuracy of the final results of Logistic Regression and Cox Model could be very high, the models failed to detect real ESRD patients among all participants. Almost all patients were predicted as non-ESRD patients,

leading to an accuracy as high as 0.981 due to the highly unbalanced data structure. In contrast, SMOTE-ENN+Logistic Regression detected 86.3% ESRD patients at the cost of misclassification of 45.2% CKD patients.

Table 4.3: Prediction performance of Logistic Regression and Cox Models

	SMOTE-ENN+Logistic Regression		Logistic Regression		Cox Model	
	Condition Positive	Condition Negative	Condition Positive	Condition Negative	Condition Positive	Condition Negative
Predicted Positive	TP=44	FP=1291	TP=2	FP=6	TP=1	FP=2
Predicted Negative	FN=7	TN=1567	FN=49	TN=2852	FN=50	TN=2856
Accuracy	0.554		0.981		0.982	
AUC	0.687		0.7		0.693	
Recall	0.863		0.039		0.02	
Precision	0.033		0.25		0.333	
F3	0.245		0.043		0.023	

Note: TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

Other evaluation measures are also presented in Table 4.3. It was not surprising that SMOTE-ENN+Logistic Regression had the lowest accuracy, 0.554, compared to the 0.981 and 0.982 of the other two models. AUC of all three models were relatively the same on the same level, making it difficult to distinguish the performances. If the aim was only to detect positive cases and Recall was used to evaluate the model without considering the cost of misclassification of negative cases, the performance of SMOTE-ENN+Logistic Regression would be incredibly good with a Recall of 0.863, compared with the values of 0.039 and 0.020 of the other two models. While the situation would be totally different if Precision was used. In this case, it was more reliable to use F_3 as the measure to evaluate the performance, as it was a combination of both Precision and Recall, with a penalty on Precision to address the importance of detecting ESRD patients accurately.

4.3.5 Threshold comparison

The evaluation metrices of the results with different thresholds are shown in Table 4.4. The results of Cox Model with threshold of 0.05 and 0.01 had significant increase in the number of positive cases, however, the growth in positive numbers was at a higher cost of misclassification of negative cases than SMOTE-ENN+Logistic Regression ($FP = 2378$ when $threshold = 0.01$). Logistic regression with threshold of 0.7 had relatively better results, $Accuracy = 0.752, F_3 = 0.300$, when compared with the results of default threshold, $Accuracy = 0.554, F_3 = 0.245$.

Table 4.4: Prediction performance of Cox Models with different thresholds

	SMOTE-ENN+Logistic Regression				Cox Model			
	RP = 0.7		RP = 0.6		SP = 0.05		SP = 0.1	
	Condition Positive	Condition Negative	Condition Positive	Condition Negative	Condition Positive	Condition Negative	Condition Positive	Condition Negative
Predicted Positive	TP=36	FP=705	TP=41	FP=987	TP=26	FP=773	TP=43	FP=2378
Predicted Negative	FN=15	TN=2153	FN=10	TN=1871	FN=25	TN=2085	FN=8	TN=480
Accuracy	0.752		0.657		0.726		0.18	
Recall	0.706		0.804		0.51		0.843	
Precision	0.049		0.04		0.033		0.018	
F3	0.3		0.276		0.207		0.149	

Note: RP = Risk Probability threshold, SP = Survival Probability threshold, TP = True Positive, FP = False Positive, TN = True Negative, FN = False Negative

4.4 Discussion

4.4.1 Imbalanced data

The imbalanced data structure caused by low prevalence of some diseases were rarely discussed in previous studies. Survival models or logistic regression were applied in the most studies, with AUC or C statistics as evaluation measures. As shown in Table 4.3, the model performance looked satisfactory only because these evaluation measures were not able to

reflect the real prediction performance. The C statistics of the Cox model in our paper was 0.807 and AUC was 0.693, which were both comparable to previous studies [143, 95, 50]. However, the confusion matrix showed that the good performance was caused by the fact that the models predicted almost all results as negative cases and it could not give reasonable results by changing the survival probability thresholds.

Our paper proposed a solution to the imbalanced structure by applying resampling method before the use of logistic regression. It was not difficult to introduce the resampling methods into the disease prediction, as it was more like a data preprocessing procedure, instead of adding complexity to the model. But the simple action solved the problem of imbalanced data structure, making the logistic regression able to detect the real ESRD patients.

One of the limitations was that 45% of negative cases were misclassified, but it might be more important to accurately detect ESRD patients than asking more patients to take good care of themselves when they were at low risk. To prevent the progression to ESRD, extra efforts include timely treatment of the underlying disease [139], anti-hypertensive therapy [60, 31], anti-diabetic therapy [103, 197], and life style management [9]. In clinical practice, the medical care to prevent ESRD does not require additional medication. The identified CKD patients with high risk of ESRD will benefit for a more intensive follow-up examination of blood pressure, glucose, and diet, which can be done at home by themselves at low costs. Therefore, it is not a heavy burden for the CKD patients identified as high-risk. Besides, the subsequent analysis can be done for the identified CKD patients after tracking them for a period. The data used in this study was extracted from the primary care data which was collected without a focus on a specific disease. Once the high-risk CKD patients are identified, with more frequent and maybe more specific examinations, it is possible to further predict the true positive cases much more accurately from all identified CKD patients.

Although the misclassification rate of negative cases seems to be high, significant improvements can be observed when compared with Cox model. To accurately identify 44 ESRD patients out of 51 using our method, 1291 CKD patients were mistakenly marked as “high risk”. However, when we used Cox Model (threshold=0.01) to identify 43 ESRD patients, 2378 out of 2858 CKD patients were marked as “high risk”. The traditional prognostic models are unable to successfully detect the true positive cases, it is safer to track all CKD patients and give them frequent examinations, that means, 1 out of 57 tracked patients will truly progress to ESRD in the future. By using our method, we narrowed the size of the tracked group and 1 out of 30 patients will be the ESRD patients in the future.

When the threshold of logistic regression was set as 0.7, the predictive performance was relatively better. About 75% patients (70.6% ESRD patients and 75.3% CKD patients) were correctly predicted. The overall performance was better but this was achieved by tolerating a lower accuracy in detecting true positive cases. In fact, it is possible to select an optimal threshold using machine learning techniques. One commonly used method is to do grid search based on a specific evaluation measures. For example, the step of the grid search can be 0.01, then every possible threshold, e.g. 0.23, 0.59, are tested and evaluated based on the predefined evaluation measures. The threshold leading to the best performance is regarded as the optimal threshold. We didn’t do the threshold parameter tuning, mainly because it is difficult to give a definite answer what is the best result. As the example of threshold of 0.7, it has better overall performance but has lower accuracy in detecting the true positive cases. We think this need further discussion among a larger groups of clinical experts to what extent the trade-off should be made. Therefore we didn’t give the optimal threshold, but gave some thresholds that are representative. The thresholds for logistic regression showed the fact that the overall prediction may be better when having a higher threshold, but the accuracy of true positive case will drop. The thresholds of survival models are to show the advantage of our method

when compared with survival models. Logistic (Threshold=0.5) and survival model (SP=0.1) have very close values in true positive cases, but survival model has almost two folds of false positive. And all of these discussion and examples prove that it is necessary to keep in mind that False Positive should be penalized and that the imbalanced data structure should be specifically dealt with.

4.4.2 Evaluation Measures

Because of the imbalanced data and the reality that False Negative and False Positive should not be treated equally, it was very difficult to reflect the true predicted results, thus, the evaluation measures should be carefully compared and selected. Most studies used accuracy and AUC as evaluation measures, and Table 4.3 reveals that they could not work well in this complex situation. Similarly, Recall and Precision could only show one aspect of the prediction. F-score was chosen as the evaluation measure in the end, with the parameter set as 3. Even a different choice of parameter could lead to completely different conclusion, as it stands for the level of penalty to False Positive. As shown in Table 4.6 in Section 4.6.2, when the parameter was 1, meaning False Negative and False Positive were treated equally, Logistic Regression without any special procedures turned out to be the model with the best performance, which was clearly not true. SMOTE-ENN+Logistic Regression started to gain benefits when the parameter was higher, adding penalty to False Positive. When it was 2, SMOTE-ENN+Logistic Regression was already the best model, but the gap in the F-scores was not large enough to convince the difference. In this case, we chose to use F_3-score as the evaluation measure in this paper.

4.4.3 Generalization

The topic of this paper is to predict ESRD patients out of CKD patients. However, the same problem of imbalanced data structure can be commonly observed for many chronic diseases. As mentioned in Introduction, it is regarded as moderate-level imbalance if the proportion of minority class is within the range 1%-20%. Few chronic diseases have a prevalence higher than 20%, instead, many chronic diseases had the prevalence within the range 1%-20%. Based on the definition of the rare diseases [36], the prevalence can be as low as 5 in 10,000. Therefore, the imbalanced data structure is a data issue that requires special attention for almost all diseases when giving prognostic prediction. The resampling technique presented in our study can be widely used for the studies on other diseases in the future. It is not necessary to use logistic regression together with SMOTE-ENN all the time, the logistic regression can be replaced with some other prognostic models. The resampling technique can be regarded as a data preprocessing procedure that alleviates the issue of imbalanced data structure, and can be easily combined with other methods. It is the same for the evaluation measure. F-score can be widely used when the False Negative and False Positive cases should not be treated equally.

4.5 Supplementary materials

4.5.1 Supplementary methods

Resampling methods

There are mainly 3 types of resampling methods, i.e. oversampling methods, undersampling methods, and a combination of over- and under-sampling methods. Oversampling is a method to duplicate samples from the minority

class. The representative methods include Random Oversampling, SMOTE, and ADASYN. All of these three methods were tested. Undersampling is a method to delete samples from the majority class. The representative methods include Random Undersampling, Near Miss (Parameter = 1, 2, or 3), Condensed Nearest Neighbor, Tomek Links, and Edited Nearest Neighbor. All of these methods were tested, including different parameters from 1 to 3 for Near Miss. A combination of over- and under-sampling method is a hybrid method which duplicate samples in the minority and delete samples from the majority. We tested SMOTE-ENN and SMOTE-Tomek.

4.5.2 Supplementary tables

Table 4.5: Full list of risk factors

Risk Factors	**Description / Details**
Age	/
Gender	/
Hemoglobin	/
Uric acid	/
Disease	**ICPC Codes**
Hypercholesterolaemia	T93
Type 2 diabetes mellitus	T90
Obesity	T82, T83
Hypertension	K85, K86, K87
Malignancy	Malignancy NOS (A79), Hodgkin's disease/lymphoma (B72), Leukaemia (B73), Malignant neoplasm blood other (B74), Benign/unspecified neoplasm blood (B75), Malignant neoplasm stomach (D74), Malignant neoplasm colon/rectum (D75), Malignant neoplasm pancreas (D76), Malignant neoplasm digest other/NOS (D77), Neoplasm digest benign/uncertain (D78), Neoplasm of eye/adnexa (F74), Neoplasm of ear (H75), Neoplasm cardiovascular (K72), Malignant neoplasm musculoskeletal (L71), Neoplasm benign/unspecified musculo. (L97), Malignant neoplasm nervous system (N74), Neoplasm nervous system unspecified (N76), Malignant neoplasm bronchus/lung (R84), Malignant neoplasm respiratory, other (R85), Neoplasm respiratory unspecified (R92), Malignant neoplasm of skin (S77), Malignant neoplasm thyroid (T71), Neoplasm endocrine other/ unspecified (T73), Malignant neoplasm of kidney (U75), Malignant neoplasm of bladder (U76), Malignant neoplasm urinary other (U77), Neoplasm urinary tract NOS (U79), Malignant neoplasm relate to pregnancy (W72), Malignant neoplasm cervix (X75), Malignant neoplasm breast female (X76), Malignant neoplasm genital other (f) (X77), Genital neoplasm other/unspecified (f) (X81), Malignant neoplasm prostate (Y77), Malignant neoplasm male genital other (Y78) and Benign/unspecified neoplasm gen. (m) (Y79)
Medication	**ATC Codes**
Antihypertensives	C02
Diuretics	C03
Beta blocking agents	C07
Calcium channel blockers	C08
Agents acting on the renin–angiotensin system	C09
Lipid modifying agents	C19

Table 4.6: Comparison of F scores with different parameters to show the effect of the choices of evaluation measures

	F1	F2	F3
SMOTE-ENN+Logistic Regression (RP=0.5)	0.064	0.143	0.245
SMOTE-ENN+Logistic Regression (RP=0.6)	0.076	0.166	0.276
SMOTE-ENN+Logistic Regression (RP=0.7)	0.091	0.19	0.3
Logistic Regression	0.068	0.047	0.043
Cox Model (SP=0.5)	0.037	0.024	0.022
Cox Model (SP=0.05)	0.061	0.13	0.082
Cox Model (SP=0.01)	0.035	0.207	0.149

Note: RP = Risk Probability threshold, SP = Survival Probability threshold

Chapter 5

Development of Multimorbidity over Time

This chapter has been published by Journals Of Gerontology Series A as:

Shi, X., Nikolic, G., Van Pottelbergh, G., van den Akker, M., Vos, R., De Moor, B. (2021). Development of multimorbidity over time: an analysis of Belgium primary care data using Markov chains and Weighted Association Rules Mining. Journals Of Gerontology Series A:76(7):1234-1241. doi: 10.1093/gerona/glaa278

ABSTRACT

Background: The prevalence of multimorbidity is increasing in recent years and patients with multimorbidity often have a decrease in quality of life and require more health care. The aim of this study was to explore the evolution of multimorbidity taking the sequence of diseases into consideration.

Methods: We used a Belgian database collected by extracting coded

parameters and over 100 chronic conditions from the Electronic Health Records of general practitioners to study patients older than 40 years with multiple diagnoses between 1991 and 2015 (N= 65,939). We applied Markov chains to estimate the probability of developing another condition in the next state after one diagnosis. The results of Weighted Association Rules Mining (WARM) allow us to show strong associations among multiple conditions.

Results: About 66.9% of the selected patients had multimorbidity. Conditions with high prevalence, such as hypertension and depressive disorder, were likely to occur after diagnosis of most conditions. Patterns in several disease groups were apparent based on the results of both Markov chain and WARM, such as musculoskeletal diseases and psychological diseases. Psychological diseases were frequently followed by irritable bowel syndrome.

Conclusion: Our study used Markov chains and WARM for the first time to provide a comprehensive view of the relations among 103 chronic conditions, taking sequential chronology into consideration. Some strong associations among specific conditions were detected and the results were consistent with current knowledge in literature, meaning the approaches were valid to be used on larger datasets, such as National Healthcare Systems or private insurers.

Key words: chronic conditions, chronology of disease, machine learning

5.1 Introduction

Multimorbidity – the co-occurrence of two or more chronic diseases [188, 189] – is frequent, especially among older people, and its occurrence is increasing. Consequences of multimorbidity include a decrease in quality of life and functional status, and an increased health care utilization [198, 102, 81].

Increasing studies look into frequent combinations of diseases, most often duals and triads [152, 80].

Ng et al. [116] made a systematic review of analytical methods used to identify multimorbidity patterns. The results showed that more than half of the studies (62/103) only adopted descriptive measures of multimorbidity to explain the associations between health conditions and selected indices, and 90% of the left 41 studies applied factor analysis or clustering. There are some other studies using methods like latent class analysis or Latent Dirichlet Allocation [122, 127]. These models are similar to clustering, which creates disease groups based on measures constructed by the algorithms. Apart from the disadvantage of being unable to track the development of multimorbidity, the results from these methods are often clusters of diseases instead of one specific disease, making it difficult to provide accurate clinical decision support in practice.

Some studies used Association Rule Mining (ARM) to explore multimorbidity [201, 71]. ARM is a method to uncover the combinations of items that occur together frequently [134, 7]. However, in traditional ARM, the sequence of the items is not taken into account. Besides, association rules among less frequent diseases will turn out to be strong if disease prevalence vary in a large range, because with fewer observations, the weight of each observation will take up a larger percentage. If two low-prevalence diseases with limited observations coincidentally co-occur for several times, it will result in strong association rules when using traditional ARM, while these cases are of noninterest. Weighted Association Rule Mining (WARM) is a good solution to overcome the two limitations of traditional ARM [141]. It introduces weights to make the co-occurrence with the same items different from each other if the sequence changes and the model adjusts the weight of the item and combination based on selected criteria and prioritizes the rules according to their importance.

Previous studies on multimorbidity mostly include a small group of common

chronic diseases. Held et al. [71] studied about 17 common diseases for the men aged 70 years or older, such as heart diseases, diabetes, and stroke, and Zemedikun et al. [201] collected data on 36 conditions. As a result, a large number of less frequent chronic conditions were neglected, excluding the potentially relevant information concerning a large number of patients, as 6-8% of the general population is affected by a rare disease [106].

Models applied in previous studies group the diseases based on associations or correlations, but are not able to present the evolution of multimorbidity over time. Analyses over time regarding patient trajectories of multimorbidity can help to identify vulnerable patient groups and provide suggestions for improving care for complex patients. This study aims to take into account the sequence of occurrence. Another contribution of this paper is that the associations among over 100 chronic conditions were included in the analyses.

5.2 Methods

5.2.1 Study Design and Participants

We used Intego [120], a longitudinal database of Electronic Health Records of patients in general practice from the Flanders region in Belgium. About 300,000 individual patients are recorded in the database, corresponding to over 2.3% of the Flemish population, and they are representative for the general population in Flanders, Belgium [184]. Intego uses an opt-out methodology and is approved by the local ethical committee of the KU Leuven and in line with Belgian privacy regulations. The register used for our analyses is based on daily clinical practice in primary care. All general practitioners (GP) participating in the Intego registry are trained and receive regular feedback on their registration skills to minimize differences

in prevalence caused by registry. Only the data of sufficient quality are used and quality of the data is secured through the testing of quality criteria[184].

Instead of the whole age span, health records of patients older than 40 were selected because multimorbidity is particularly prevalent in adults aged 40 years and older. For this study we analyzed the data concerning the period between1991 and 2015. A total of 98,632 patients fulfilled the requirements and had records of the selected chronic conditions. The study regarded it as multimorbidity if the patient had more than one out of the selected chronic conditions recorded in an overlapping period. As these health conditions were all chronic, it was assumed that they lasted until the end of follow-up. In total, 65,939 patients had multimorbidity and among this group of patients, the average duration between the first and the last diagnosis was 8.29 years, the median value is 7.33 (P25-P75: 2.79-12.87). The sequence of the health conditions were derived from the diagnosis dates. The distribution of the total patients and multimorbidity patients in each year is shown in Table 5.3 in Section 5.5.2.

Intego contains all the coded data registered in general practices, including clinical parameters, laboratory tests, disease diagnosis and prescriptions. The diagnosis are coded based on the International Classification of Primary Care -2 (ICPC-2) and the exact date of the diagnosis is also coded [94]. ICPC is a classification system aiming to reflect the content of primary care, with 17 chapters for different disease groups and about 1,300 codes. ICPC codes cover the most frequent complaints or symptoms that GPs encounter in primary care and allow classification of the patient’ s reason for encounter, the diagnosis managed, interventions, and the ordering of these data in an episode of care structure. Therefore, ICPC codes include not only medical diagnosis, but complaints, lifestyle factors, and risk factors, for example alcohol abuse and limited function.

5.2.2 Selection of Conditions

A dedicated group of 4 GPs and an epidemiologist of the Academic Centre for General Practice of the KU Leuven, categorized all diagnostic codes and complaints codes of the ICPC, into either acute or chronic, adjusted from a previous categorization of ICPC codes [117]. The purpose of this was to come to more valid prevalence estimates of chronic conditions and was based on clinical experience and epidemiological knowledge concerning the duration of a certain condition. An expected duration of 3 years or longer was classified as chronic. All five members of the group prepared the categorization independently, and the final categorization was discussed until consensus in 4 group meetings. This resulted in a selection of 105 conditions that were considered chronic. However, there was no record of A21 (Risk factor for malignancy) in Intego and we only had one record of W72 (Malignant neoplasm related to pregnancy), which had no co-occurrence with any other condition. As a consequence, A21 and W72 were excluded from analysis, as the following analytics required patients with at least two conditions in the dataset. The list of the 103 chronic conditions can be found in Table 5.4 in Section 5.5.2.

5.2.3 Statistical Analyses

The prevalence of multimorbidity in the selected population was estimated by sex and age groups [201]. Then Markov chains were applied to study the sequence of development of selected conditions. A Markov chain is a stochastic model that describes a sequence of possible events whose probability only depends on the state attained in the previous event [63]. The result of this model is the transition matrix, whose elements can be interpreted as the probability of having condition B after condition A. In this study, we calculated the first-order transition matrix, meaning the

transfer from condition A to condition B occurs in only one step. More details on Markov chains can be found in Section 5.5.1.

The first-order transition matrix estimates the probability of transition between two conditions. To further select closely related groups of conditions, Weighted Association Rules Mining (WARM) was applied [141]. In WARM algorithm, the record of one condition is called an item and the co-occurrence of several conditions is called a transaction. For example, $Hypertension, Diabetes, Stroke$ is a transaction with three items. The results of WARM are rules derived from a large amount of transactions and can be written as $X \rightarrow Y$, X and Y are two different sets of items, known as itemsets. The itemsets can include either one or multiple conditions and they can differ from the combinations appeared in the transactions, as these itemsets are the summarized results instead of original data. Using the example transaction together with some other transactions, it is possible to derive a rule $Diabetes \rightarrow Hypertension$.

Three important measures can be used to evaluate the strength of association rules, namely support, confidence and lift. Support refers to the frequency of the itemset and higher support indicates the rules have high appearance among all the transactions. Confidence indicates how often the rules are true and it reveals reliability of the rule. Lift is the ratio to define the importance of the rule. A lift larger than one means the itemset Y is likely to occur if itemset X occurs and larger value indicates closer associations. Thresholds for each criterion were defined to select the strong rules. Based on the distribution of all the measures, the selection criteria in the paper were set as $support > 0.001, confidence > 0.1$, and $lift > 1$. The calculation of weighted association rules was done with R package arules [76]. More details on WARM can be found in Section 5.5.1.

5.3 Results

5.3.1 Multimorbidity and Prevalence

About two thirds of the selected patients had multimorbidity (Table 5.1). The prevalence was much higher than that of the general population, mainly because the majority of the selected patients were older than 60 years. Based on the age in 2015, 63.7% of the patients were in the group of 60 years old or above. Females had a higher prevalence than males and the rates were evidently higher in the elder age groups. Among patients over age 74, 77.8% had multimorbidity.

Table 5.1: Prevalence of multimorbidity by gender group and age group

	Number of Patients in Intego (N (%))	Number of Multimorbidity Patients (N (%))	Prevalence rate of multimorbidity (%)
	TN	MN	MN/TN
Total	98632	65939	66.9
Gender group			
Male	45796 (46.4)	29712 (45.1)	64.9
Female	52836 (53.6)	36227 (54.9)	68.6
Age group in 2015			
40-49	13400 (13.6)	6719 (10.2)	50.1
50-59	22412 (22.7)	13017 (19.7)	58.1
60-74	29764 (30.2)	20482 (31.1)	68.8
$\geqslant 75$	33056 (33.5)	25721 (39.0)	77.8

The prevalence of each chronic condition is shown in Table 5.4 in Section 5.5.2. Top 5 conditions with the highest prevalence were hypertension, lipid disorder, depressive disorder, type 2 diabetes and osteoarthrosis. In contrast to these common conditions, some conditions only have limited records, such as poliomyelitis and neoplasm urinary tract. In total, 12 chronic conditions had less than 100 observations while 17 chronic conditions had a prevalence

higher than 5%. The prevalence of the other 35 chronic conditions ranged from 1% to 5%.

5.3.2 Probabilities of Follow-Up Conditions

Figure 5.1 is a heatmap of the transition probability matrix calculated by using a Markov chain. The conditions distributed on the y-axis are the conditions at the first status, and those on the x-axis stand for the second status. Each small box in the heatmap represents the probability of developing the condition on the x-axis after the diagnosis of the condition on the y-axis and darker color means higher probability. The scale of the color bar on the right side of the heatmap was defined between 0 and 0.1 because most of the probabilities were within this range. The probabilities under 0.1 could be distinguished by the color shade while the probabilities higher than 0.1 would have the same darkest color as the probability of 0.1.

Some patterns can be observed from the heatmap and can be explained based on the properties of these chronic conditions. Distinct columns with deep color often appear when the conditions on the x-axis have high prevalence, for example, hypertension (K86), depressive disorder (P76), diabetes (T90) and lipid disorder (T93). There are also some rows that are totally white in most of the boxes, making a visual horizontal division. These rows start from conditions with low prevalence on the y-axis. They have fewer records, leading to a higher probability in each observed co-occurrence and a probability of 0 if no co-occurrence was observed. Poliomyelitis (N70), neoplasm cardiovascular (K72) and limited function/disability coming after psychological disorder (P28) are representatives of these conditions. Apart from the columns and rows, there are some triangle areas with deep color, including multiple conditions from the same disease group. Take cardiovascular diseases as an example, K74-K86 form the largest triangle area, meaning that once the patients develop one condition out of the

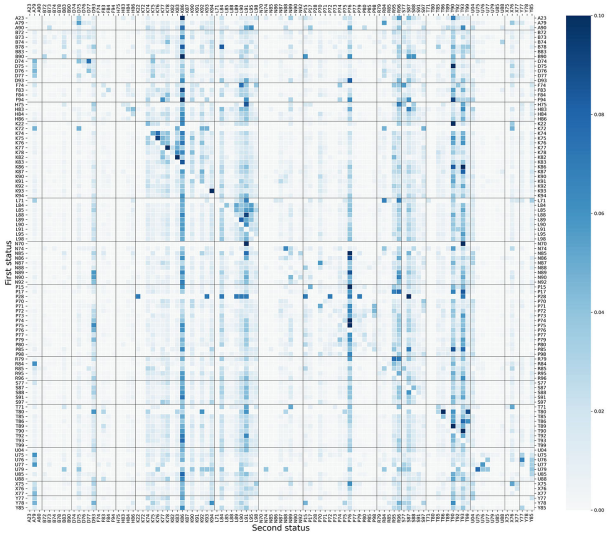


Figure 5.1: Heatmap of the first-order transition probability matrix of a Markov chain including 103 chronic conditions. The number in each square is the probability of developing the condition on the x-axis after the diagnosis of the condition on the y-axis. The scale of the color bar was defined between 0 and 0.1. Vertical and horizontal lines were added for disease groups. (Disease groups: A-General, B-Blood and Immune, D-Digestive, F-Eye, H-Ear, K-Cardiovascular, L-Musculoskeletal, N-Neurological, P-Psychological, R-Respiratory, S-Skin, T-Metabolic, U-Urological, X-Female Genital, Y-Male Genital)

group, they are at higher risk for other cardiovascular diseases. Some other disease groups, such as musculoskeletal diseases, endocrine, metabolic and nutritional diseases and blood diseases, have similar patterns. Almost all these triangles are above the diagonal, because in general the conditions from the same group are more severe if they have larger number in ICPC codes, thus, it is more probable that they occur in the late stage as a subsequent condition.

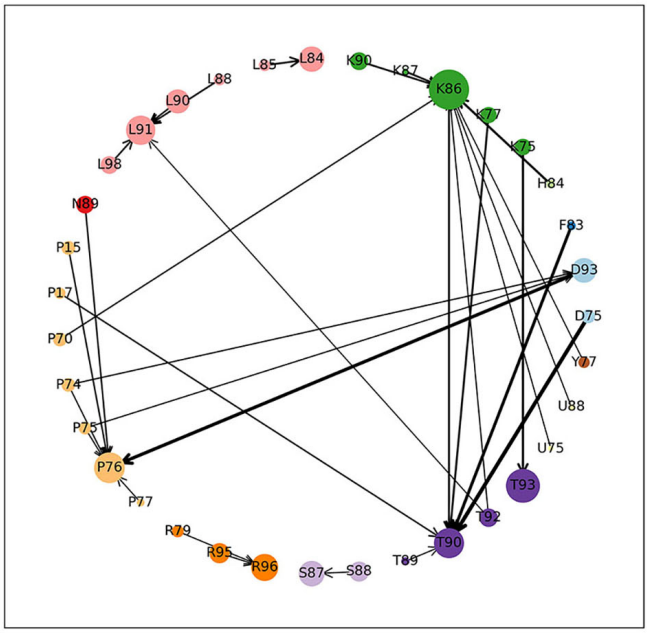
The heatmap can also be interpreted in terms of a specific condition. For instance, the cells with dark color shade in the row of chronic alcohol abuse (P15, prevalence=2.915%), a complaint with a medium level of prevalence in the dataset, showed that it had a higher probability to have K86, P76, T90 and T93 as its subsequent. It was also likely to develop chronic obstructive pulmonary diseases (R95), asthma (R96), and dermatitis/atopic eczema (S87) afterwards. The dark blue box representing the probability of developing depressive disorder (P76) after chronic alcohol abuse requires special attention, implying a strong association between them. On the other hand, when observing the conditions prior to chronic alcohol abuse, HIV-infection/aids (B90), limited function/disability (P28), malignant neoplasm male genital other (Y78) and epilepsy (N88) became conspicuous. In this way, the clinical interested investigators can inspect the heatmap for clinically relevant associations of their interest.

As it may be difficult to go through the whole heatmap to find all interesting associations, we conducted an automatic search to detect the important associations. The method is described in details in Section 5.5.1 and the results are shown in Table 5.5 in Section 5.5.2.

5.3.3 Weighted Association Rules

The results of the Markov chain analysis included 103 chronic conditions to show a general picture of condition relations. Then Weighted Association

Rules Mining (WARM) was applied to calculate strong associations from all chronic conditions and pay attention to robust and strong relations instead of incidental co-occurrence. Table 5.2 shows the results of WARM, which are the strongest rules selected based on the thresholds of support, confidence and lift. These rules were presented in a more intuitive way in Figure 5.2. The results present the sequence of the condition diagnoses that occurs more often than other possibilities based on the statistics, however, it does not mean causality. For example, when the antecedents is "Retinopathy" and the consequents is "Diabetes", it shows the fact that one of the early signs of diabetes is blurred vision so that the patients are diagnosed as diabetes after they have checked their eyes, but in fact retinopathy is the complication of diabetes. In some cases, we observed strong relations in both directions, for example, depressive disorder and irritable bowel syndrome, which meant these two conditions were probably strongly correlated and they often occurred together, making the sequence indifferent. Similar to the findings from Figure 5.1, hypertension and diabetes were subsequent of many chronic conditions and many strong relations were among the same disease group, such as musculoskeletal diseases, psychological diseases, respiratory diseases and skin diseases. But there were still some rules that were worth special attention. Severe pain, such as migraine, could be followed by depressive disorder, while multiple psychological diseases might be followed by irritable bowel syndrome. As was already known in some studies, tobacco abuse could precede diabetes, which was also reflected in the WARM results. Besides, associations among dementia and hypertension, gout and osteoarthrosis, and presbyacusis and hypertension were also detected.



D75 - Colon cancer
D93 - Irritable bowel syndrome
F83 - Retinopathy
H84 - Presbyacusis
K75 - Myocardial infarction
K77 - Heart failure
K86 - Hypertension uncomplicated
K87 - Hypertension complicated
K90 - Stroke
L84 - Back pain
L85 - Acquired deformity of spine
L88 - Rheumatoid
L90 - Osteoarthrosis of knee
L91 - Osteoarthrosis other
L98 - Acquired deformity of limb
N89 - Migraine
P15 - Chronic alcohol abuse
P17 - Tobacco abuse
P70 - Dementia
P74 - Anxiety disorder
P75 - Somatization disorder
P76 - Depressive disorder
P77 - Suicide/suicide attempt
R79 - Chronic bronchitis
R95 - Obstructive pulmonary dis
R96 - Asthma
S87 - Dermatitis
S88 - Dermatitis contact
T89 - Diabetes insulin dependent
T90 - Diabetes non-insulin dependent
T92 - Gout
T93 - Lipid disorder
U75 - Kidney cancer
U88 - Glomerulonephritis
Y77 - Prostate cancer

Figure 5.2: Visualization of Weighted Association Rules Mining results. The color of the nodes stands for the disease group and the size of the nodes is determined by the frequency of the occurrence of the condition. The arrows on the edge show the direction of the relation and wider edges imply that the associations were closer and stronger.

Table 5.2: Results of Weighted Association Rules Mining

Antecedents	Consequents	Support	Confidence	Lift
Suicide/suicide attempt	Depressive disorder	0.00139	0.505	3.384
Retinopathy	Diabetes non-insulin dependent	0.00236	0.521	2.923
Retinopathy AND Hypertension	Diabetes non-insulin dependent	0.00129	0.476	2.67
Anxiety disorder/anxiety state	Depressive disorder	0.00373	0.297	1.995
Acquired deformity of spine	Back syndrome w/o radiating pain	0.00106	0.12	1.986
Somatization disorder	Depressive disorder	0.00268	0.264	1.767
Somatization disorder	Irritable bowel syndrome	0.00136	0.134	1.738
Rheumatoid/seropositive arthritis	Osteoarthrosis other	0.00207	0.193	1.697
Dermatitis contact/allergic	Dermatitis/atopic eczema	0.00521	0.136	1.647
Diabetes insulin dependent	Diabetes non-insulin dependent	0.00155	0.292	1.64
Chronic alcohol abuse	Depressive disorder	0.00405	0.231	1.551
Chronic bronchitis	Asthma	0.00132	0.153	1.513
Migraine	Depressive disorder	0.00812	0.221	1.484
Depressive disorder	Irritable bowel syndrome	0.01565	0.105	1.36
Irritable bowel syndrome	Depressive disorder	0.01565	0.203	1.36
Anxiety disorder/anxiety state	Irritable bowel syndrome	0.00131	0.105	1.356
Hypertension complicated	Hypertension uncomplicated	0.00207	0.782	1.266
Malignant neoplasm of kidney	Hypertension uncomplicated	0.00105	0.705	1.141
Tobacco abuse	Diabetes non-insulin dependent	0.00173	0.202	1.133
Gout	Osteoarthrosis other	0.00537	0.127	1.114
Heart failure	Diabetes non-insulin dependent	0.00171	0.195	1.092
Presbyacusis	Hypertension uncomplicated	0.0023	0.668	1.082
Chronic obstructive pulmonary disease	Asthma	0.00365	0.109	1.08
Acute myocardial infarction	Lipid disorder	0.00587	0.304	1.046
Gout	Hypertension uncomplicated	0.02732	0.645	1.044
Glomerulonephritis/nephrosis	Hypertension uncomplicated	0.00121	0.641	1.038
Malignant neoplasm colon/rectum	Diabetes non-insulin dependent	0.00281	0.182	1.021
Osteoarthrosis of knee	Osteoarthrosis other	0.00634	0.115	1.015
Stroke/cerebrovascular accident	Hypertension uncomplicated	0.01438	0.627	1.015
Malignant neoplasm prostate	Hypertension uncomplicated	0.00636	0.626	1.013
Dementia	Hypertension uncomplicated	0.00742	0.625	1.013
Acquired deformity of limb	Osteoarthrosis other	0.00287	0.114	1.005

5.4 Discussion

In this study, we found that about 67% of the selected patients had multimorbidity and the rate could be as high as 77.8% among patients older than 74 years old. Conditions with relatively high prevalence, such as hypertension, depressive disorder, diabetes and lipid disorder, were very likely to occur after diagnosis of most conditions as they are the common complications of many conditions. Multiple conditions from the same disease group often happened together, for example once the patient developed one condition from the cardiovascular disease group, he would be at higher risk of other cardiovascular diseases. Several morbidity clusters were apparent based on the results of the Markov chain and WARM, such as musculoskeletal diseases, psychological diseases, respiratory diseases and skin diseases. Hypertension and diabetes were subsequent of many chronic conditions and multiple psychological diseases might be followed by irritable bowel syndrome afterwards.

5.4.1 Strengths and Limitations

In this paper, the Markov chain analysis and WARM were used to study the multimorbidity among people aged over 40 years old, making it one of the first to investigate multimorbidity considering the sequence of evolution. Besides, instead of using surveys and self-reported health complaints as previous studies did, our paper used the registered chronic health conditions from Belgian data collected from patients' daily consultation with their GP. The real medical records made our study reliable and generalizable to the Belgian population. Moreover, we presented a comprehensive view of the associations among 103 chronic conditions, covering chronic conditions from all disease groups, which might shed light to some potential relations that were not widely discussed before. It is worth noticing that both Markov chains and WARM have rarely been applied in epidemiologic

studies, although they are very useful techniques to explore sequential rules. To our knowledge, two papers have used ARM to summarize multimorbidity patterns [127, 201], but no studies have applied WARM so far. The current results are in agreement with previous clinical findings, meaning that these approaches were valid and could be used on larger datasets in future studies, such as National Healthcare Systems or large private insurers. Because of the intensive information stored in the results of the Markov chain and the WARM, it is worth examining the results more carefully to discover potential new clinically relevant relations that have not been widely discussed before.

One limitation of the Markov chain analysis is that our results were derived under the first-order Markov assumption, meaning that the future state only depends on the current state, instead of on the sequence of history events. One example of this property is a patient with a sequence of records $Hypertension, Diabetes, Stroke, Stroke$ is State3 which only depends on State2 *Diabetes* while *Hypertension* cannot be taken into account. In this case, the transition matrix can only present the probability among two conditions. The limitation can be overcome by higher-order Markov chain , for instance, second-order Markov chain involves 3 states. However, our study included 103 chronic conditions and it was too complicated to interpret the results for all combinations in multiple steps. There are already other studies discussing about relations among disease groups or a small number of chronic conditions. Our paper would like to do a comprehensive study on the relations among over 100 chronic conditions to provide clinical decision support for the general practitioners in primary care. It is much easier to track the follow-up of one specific condition by using the Markov chains. Other methods could provide the results of disease groups (i.e. clustering and LDA) or selected results with the best criteria (i.e. WARM), but fail to provide an accurate estimation between any two individual conditions. Considering that 54.5% of the patients with multimorbidity only had 2 conditions, we would like to apply the first-order Markov chain analysis in spite of the limitation. We also used WARM in the second step which could

include multiple conditions in the itemsets. Therefore, these two methods were complementary.

Another limitation of the study is the lack of a time concept in the analysis. Although the sequence of occurrence is considered, it is not possible to incorporate exact interval time between diagnoses. Finally, the results of the Markov chain and WARM could suggest strong associations, but causal relationships remain unclear and further studies are needed to confirm causality. Thus, more studies using new methods are still expected in the future work to overcome these limitations.

5.4.2 Validity

The results of a Markov chain analysis may be questioned because the majority of the probabilities are very small, lower than 0.1. Although the values are small, they are significantly larger than the average level if we take the total number into consideration, meaning that the methodology provides significant contrast in the results to distinguish likely or unlikely sequence of diseases. It is the same in the case of WARM where the support of the selected rules might be interpreted at first sight as very low but in fact selects the strongest rules. The detailed calculation is in Section 5.5.1.

The clinical relevance was also checked. For instance, the triangle of cardiovascular diseases detected in the heatmap of Markov chain analysis reveals the fact that patients with multiple cardiovascular conditions are common in real-world clinical practice [87]. Chronic alcohol abuse was used as an example to demonstrate the Markov chain results and it could be the precursor of hypertension [157], depression [23], and diabetes [77], which was in correspondence with previous studies. Similarly, WARM selected the strongest association rules with robustness and validity among all possible combinations, hence, it is not surprising to find that almost all the association rules have been discussed in previous research, for example

the relation between psychological symptoms and irritable bowel syndrome [39, 183], tobacco abuse and diabetes [68], hypertension and many diseases such as dementia [115, 58],and the sequence of condition development was in agreement with these studies. The correspondence proves the validity of the results and shows the possibility of using the model to detect more potential relations. Considering the huge amount of combinations, it was acceptable to further lower the thresholds for support and confidence to take a look at a broader range of association rules, which might result in some potential rules that were not discussed before. The automatic search results from the Markov chain work in the same way. It provides a comprehensive view of the relations among 103 chronic conditions, making it helpful to provide decision support for practice-oriented healthcare professionals.

Finally, subgroup analyses by sex and age were conducted separately to check the robustness and consistency (Section 5.5.3). The results of Markov chain were robust in all subgroups. The horizontal and vertical belt patterns and triangle patterns were similar in all figures and the main follow-up conditions of one specific condition were almost always the same. There were some differences in WARM results among all subgroups, mainly because WARM selected the strongest and most robust rules and the importance of the rules may vary in different age and gender group [11]. Generally, more rules were detected in the group with higher prevalence of multimorbidity, but most of the rules selected in the subgroup analysis were the same as the rules in Figure 5.2, e.g. depression and irritable bowel syndrome, hypertension and lipid disorder. Some group-specific findings are in correspondence with clinical knowledge, for instance, gout did not appear in female group and younger age group, as the risk of gout is higher for the male and higher age [43, 145, 30]. Therefore, WARM is a consistent tool for subgroup analysis, and it is useful to discover group-specific findings, which can be used for subject matter experts to further examine clinical relevance.

5.5 Supplementary materials

5.5.1 Supplementary methods

Markov chains

A Markov chain is a stochastic model which describes a sequence of possible events whose probability only depends on the state attained in the previous event [63]. The "memorylessness" property of the Markov chain allows us to predict the next state of the process only based on the current state and the accuracy of the prediction will be as good as the prediction based on the full history. This stochastic model is widely used in various areas such as control system, finance, queueing problems, game theory and so on [107]. It is also applied in the area of healthcare and medical decision making, for example, to estimate transition between different status after surgery [161]. It is a useful tool to conduct cost effectiveness analysis and help decision makers learn more about possible transitions between different statuses.

Markov chains are often described by the transition matrix from time n to time $n+1$. The transition matrix describes the probabilities of particular transitions, each element of the matrix is the transition probability from state i to state j in one time step. A sequence of directed graphs are often used to represent the Markov chains and transition matrices, where the edges of graph n are labeled by the probability of going from one state at time n to the other states at time $n+1$.

Figure 5.3 shows an example of the transition matrix of a simple Markov chain and the visualization of the transition matrix. This can be regarded as estimated probability matrix of 3 diseases. The probability of the development from disease A to disease B is 0.3 (the number in the blue square in the matrix and plot) while the probability of the development from disease B to disease A is 0.4 (the number in the red square).

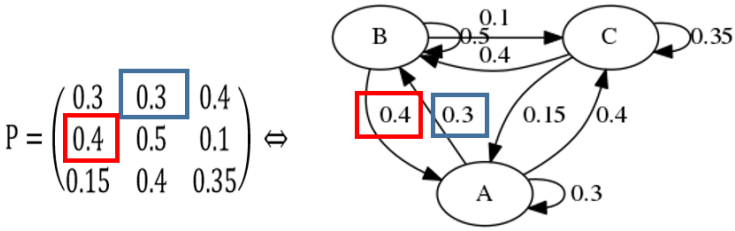


Figure 5.3: An Example of a Markov Chain Transition Matrix and Visualization

Other probabilities can be interpreted in the same way. This is a good method for the practice-oriented researchers to study the pattern of disease evolution quantitatively. An example to demonstrate how to implement the Markov chain to study multimorbidity is uploaded to Github (https://github.com/XiShi6/Multimorbidity).

The application of Weighted Association Rules Mining

Association Rule Mining (ARM) is a method to uncover the combinations of items that occur together frequently, discovering interesting relations. It is often applied in the studies of supermarket sales and customer behaviors and is gaining attention in the clinical research. There are already some studies applying ARM to discover the frequent combination of disease occurrence [201, 71]. However, as stated in Introduction of the paper, ARM doesn' t take the sequence into account. Besides, association rules among less frequent diseases will turn out to be strong if diseases with a large difference in prevalence are involved, because with fewer observations, the weight of each observation will take up a larger percentage. If two low-prevalence diseases with limited observations coincidentally co-occur for several times, it will result in strong association rules when using traditional Association Rule Mining, while these cases are of noninterest. Weighted Association

Rule Mining (WARM) is a good solution to overcome the two limitations of traditional Association Rule Mining.

One limitation of the first-order Markov chain has been fully discussed in Discussion. Therefore, WARM was applied as a supplementary method of the Markov chain. The results of WARM are rules derived from a large amount of transactions and can be written as $X \rightarrow Y$, X and Y are two different sets of items, known as itemsets. The itemsets can include either one or multiple diseases and they can differ from the combinations appeared in the transactions, as these itemsets are the summarized results instead of original data. As shown in Table 5.2, it was possible to derive a rule with multiple diseases in the itemset, $Retinopathy, Hypertension \rightarrow Diabetesnon-insulindepedent$. In this case, with a combination application of the Markov chain and WARM, we could example the patterns of diseases without the limitation of number of diseases.

WARM introduces weights to make the co-occurrence with the same items different from each other if the sequence is different. In our study, the transaction and itemset weights were calculated using the HITS algorithm [192, 174], and the frequent itemsets were found based on the Eclat algorithm [200]. The weighted support of an itemset was a generalization of support taking the weights of the transactions into consideration and the itemset was frequent if the weighted support was higher than the threshold. These weights were determined by the global link structure of the database, making it more reasonable than counting-based measurement. The weighted confidence was calculated from the weighted support of the itemset. These weighted measures can be different if the sequence changes. Although $X \rightarrow Y$ and $Y \rightarrow X$ have the same itemsets, the rule with better weighted measures means the sequence of this rule is more often observed. The weighted association rules were created from the weighted frequent itemsets following the efficient algorithm proposed by Pei et.al [128].

An example to demonstrate how to implement WARM to study multimorbidity is uploaded to Github (https://github.com/XiShi6/Multimorbidity).

Automatic search of important associations from the results of the Markov chain analysis

As it may be difficult to go through the whole heatmap to find all interesting associations, we conducted an automatic search to detect the important associations by selecting the top three diseases with the highest probability in each column and selecting the diseases with probability higher than 0.05 in each row (Table 5.5). The total sum of probability in each row equals to 1, meaning that larger probability takes up a larger share compared to all other diseases. It is possible that one disease has relatively equal probabilities around 0.0097 with all other diseases, in this case, it cannot be regarded as strong association even though the probability is the highest among the whole row. Thus, the search by row was determined by the absolute values, indicating the most likely follow-up diseases after the diagnosis of one chronic disease. In contrast, high absolute values are not very meaningful when searching by column, as columns for diseases with high prevalence have high absolute values in almost the whole column, such as the vertical line for Hypertension and Diabetes in Figure 5.1. Therefore, the top three diseases with the highest probability in the column were selected as the most likely antecedent diseases.

There is also a user-friendly way of interacting with a heatmap: that is, it is in principle possible to click on cells or regions in the heatmap to generate a more detailed picture of the pattern of relationships of the specific diseases involved; another strategy is to depict the heatmap into a graph and then visual bits and pieces of the graph. We do not present such features in this paper, but this can be done in principle for practice-oriented health care professionals or practice-oriented researchers when it is of interest to focus

on specific combinations of diseases or specific clusters of diseases.

Discussion of validity

The results of a Markov chain analysis may be questioned because the majority of the probabilities are very small, lower than 0.1. Although the values are small, they are significantly larger than the average level if we take the total number into consideration. If the probability of all other conditions diseases becoming the subsequent of one specific condition (e.g. alcohol abuse) had been equal, the average probability would have been 0.0098 (1/102). The gap is indubitable when we compare the average probability with the actual probability of developing depression (p=0.102), hypertension (p=0.060) and diabetes (p=0.044) after having diagnosis of alcohol abuse, as the actual probabilities of diabetes were 4.5 times higher than the average probability and that of depression were as large as more than 10 times. Besides, some conditions had probabilities much lower than the average probability or even close to 0. Under the principle of closer relationship leading to higher probability and the opposite with lower probability, we can observe a clear division to indicate common multimorbidity. In this case, although the values seem low, the methodology provides significant contrast in the results to distinguish likely or unlikely sequence of conditions.

It is the same in the case of WARM where the support of the selected rules might be interpreted at first sight as very low. The average probability of certain combinations is trivial when the total number is 103. The total number of possible combinations of selecting 2 out of 103 is 5253 and it is 176851 when selecting 3 out of 103. WARM does not have restrictions on how many conditions to include in one rule, leading to a huge total possible number (1.014E+31). Therefore, selected rules can be considered relevant when the selection criteria includes support > 0.001

5.5.2 Supplementary tables

Table 5.3: Distribution of the total patients and multimorbidity patients by year

Year	Total Patients			Multimorbidity Patients		
	Number Of the First Diagnosis	Number Of the Last Diagnosis	Number of Patients	Number Of the First Diagnosis	Number Of the Last Diagnosis	Number of Patients
1991	6704	1163	6704	5751	210	5751
1992	5430	1295	10971	4526	391	10067
1993	4537	1201	14213	3757	421	13433
1994	4634	1437	17646	3801	604	16813
1995	4338	1601	20547	3420	683	19629
1996	4163	1762	23109	3242	841	22188
1997	3748	1709	25095	2952	913	24299
1998	4034	2048	27420	3086	1100	26472
1999	3825	2315	29197	2816	1306	28188
2000	4952	2949	31834	3683	1680	30565
2001	3716	2663	32601	2623	1570	31508
2002	4159	3096	34097	2878	1815	32816
2003	4576	3655	35577	3119	2198	34120
2004	4295	4042	36217	2819	2566	34741
2005	4003	4101	36178	2551	2649	34726
2006	3729	4278	35806	2239	2788	34316
2007	3451	4623	34979	1982	3154	33510
2008	3416	4977	33772	1933	3494	32289
2009	3456	5436	32251	1813	3793	30608
2010	3354	6146	30169	1598	4390	28413
2011	3605	7212	27628	1578	5185	25601
2012	3097	7334	23513	1318	5555	21734
2013	2633	6763	18812	1059	5189	17238
2014	2569	7773	14618	890	6094	12939
2015	2208	9053	9053	505	7350	7350

Table 5.4: Selection of 103 chronic diseases with their prevalence in Intego

ICPC	Disease Name	Prevalence	ICPC	Disease Name	Prevalence
A23	Risk factor NOS	0.115	N88	Epilepsy	1.344
A79	Malignancy NOS	2.277	N89	Migraine	5.247
A90	Congenital anomaly OS/multiple	0.104	N90	Cluster headache	0.47
B72	Hodgkin's disease/lymphoma	0.509	N92	Trigeminal neuralgia	0.901
B73	Leukaemia	0.437	P15	Chronic alcohol abuse	2.915
B74	Malignant neoplasm blood other	0.22	P17	Tobacco abuse	1.794
B78	Hereditary haemolytic anaemia	0.121	P28	Limited function/disability (p)	0.014
B83	Purpura/coagulation defect	1.033	P70	Dementia	2.734
B90	HIV-infection/aids	0.103	P71	Organic psychosis other	0.737
D74	Malignant neoplasm stomach	0.305	P72	Schizophrenia	0.577
D75	Malignant neoplasm colon/rectum	2.748	P73	Affective psychosis	0.414
D76	Malignant neoplasm pancreas	0.363	P74	Anxiety disorder/anxiety state	2.629
D77	Malignant neoplasm digest other/NOS	0.791	P75	Somatization disorder	2.214
D93	Irritable bowel syndrome	10.145	P76	Depressive disorder	16.239
F74	Neoplasm of eye/adnexa	0.057	P77	Suicide/suicide attempt	0.72
F83	Retinopathy	0.943	P79	Phobia/compulsive disorder	0.825
F84	Macular degeneration	0.627	P80	Personality disorder	0.464
F94	Blindness	0.035	P85	Mental retardation	0.136
H75	Neoplasm of ear	0.064	P98	Psychosis NOS/other	0.632
H83	Otosclerosis	0.192	R79	Chronic bronchitis	2.048
H84	Presbyacusis	0.881	R84	Malignant neoplasm bronchus/lung	1.254
H86	Deafness	1.308	R85	Malinant neoplasm respiratory, other	0.225
K22	Risk factor cardiovascular disease	0.348	R95	Chronic obstructive pulmonary dis	6.288
K72	Neoplasm cardiovascular	0.022	R96	Asthma	12.264
K74	Ischaemic heart disease w. angina	4.955	S77	Malignant neoplasm of skin	2.512
K75	Acute myocardial infarction	3.946	S87	Dermatitis/atopic eczema	10.716
K76	Ischaemic heart disease w/o angina	3.064	S88	Dermatitis contact/allergic	6.318
K77	Heart failure	3.674	S91	Psoriasis	3.036
K78	Atrial fibrillation/flutter	6.03	S97	Chronic ulcer skin	2.402
K82	Pulmonary heart disease	0.536	T71	Malignant neoplasm thyroid	0.105
K83	Heart valve disease NOS	3.514	T80	Congenital anom endocrine/metab.	0.043
K86	Hypertension uncomplicated	2.731	T85	Hyperthyroidism/thyrotoxicosis	2.374
K87	Hypertension complicated	0.389	T86	Hypothyroidism/myxoedema	2.549
K90	Stroke/cerebrovascular accident	4.632	T89	Diabetes insulin dependent	0.92
K91	Cerebrovascular disease	0.387	T90	Diabetes non-insulin dependent	14.513
K92	Atherosclerosis/PVD	4.527	T92	Gout	5.216
K93	Pulmonary embolism	1.176	T93	Lipid disorder	18.515
K94	Phlebitis/thrombophlebitis	6.187	T99	Endocrine/metab/nutrit. dis. other	3.88
L71	Malignant neoplasm musculoskeletal	0.063	U04	Incontinence urine	4.291
L84	Back syndrome w/o radiating pain	10.059	U75	Malignant neoplasm of kidney	0.307
L85	Acquired deformity of spine	1.946	U76	Malignant neoplasm of bladder	0.485
L88	Rheumatoid/seropositive arthritis	1.926	U77	Malignant neoplasm urinary other	0.134
L89	Osteoarthrosis of hip	4.439	U79	Neoplasm urinary tract NOS	0.027
L90	Osteoarthrosis of knee	8.842	U85	Congenital anomaly urinary tract	0.304
L91	Osteoarthrosis other	13.724	U88	Glomerulonephritis/nephrosis	0.415
L95	Osteoporosis	5.424	X75	Malignant neoplasm cervix	0.264
L98	Acquired deformity of limb	4.78	X76	Malignant neoplasm breast female	3.129
N70	Poliomyelitis	0.003	X77	Malignant neoplasm genital other (f)	0.584
N74	Malignant neoplasm nervous	0.216	Y77	Malignant neoplasm prostate	1.776
N85	Congenital anomaly neurological	0.081	Y78	Malign neoplasm male genital other	0.076
N86	Multiple sclerosis	0.277	Y85	Benign prostatic hypertrophy	3.174
N87	Parkinsonism	1.452			

Table 5.5: The important associations from the Markov Chain results

Antecedents	Disease	Consequents	Antecedents	Disease	Consequents
K22 P72 P80	A23	K86 R96 S87 T90	N74 T80 U79	N88	K86
R84 U75 U77	A79	D75 R84	A23 N90 T71	N89	D93 K86 L91 P76 R96 T93
H84 P98 R85	A90	K86 L91 L95 P76 T93	N74 N89 N92	N90	D93 K86 L91 N89 P76 R96 T93
B73 B74 B90	B72		A23 N85 N90	N92	D93 K86 L91 P76 R96
A90 B74 U77	B73	K86	B90 P28 Y78	P15	K86 P17 P76 T90 T93
A79 B73 N74	B74	K78	A90 N90 P15	P17	P76 R95 R96 T90 T93
B72 N92 P98	B78	K86 L84 T90	L90 N88 N89	P28	K22 K93 L84 L89 L90 L91 P15 P72 P76 P79 S87 T93
B78 B90 U79	B83	K86	K90 N87 P71	P70	
B72 U85 X75	B90	K86 P15 P76 S87 S88 T90	K91 N87 P70	P71	P70 P76 P98
A79 B72 H83	D74	A79 D75 D77 K75	N86 P28 P98	P72	P76 P98
A79 D74 K72	D75	T90	B78 P72 P80	P73	P76 T93
A79 B73 D74	D76	A79	F74 F94 P79	P74	D93 P76 T93
D74 D76 R85	D77	K86 T90	P74 T80 Y78	P75	D93 K86 P76 T93
N89 P75 T80	D93	K86 L91 P76 R96 S87 T93	N85 P74 P75	P76	D93 K86 T93
H86 K93 U77	F74	L90 S77	P73 P76 P80	P77	P76
T80 T89 U88	F83	K86 T90 T93	N85 P28 P74	P79	D93 K86 P76 T93
B78 F83 H84	F84	K78 K86	A90 P73 P77	P80	K86 P76 T93
F83 K87 P98	F94	K77 K86 L91 P76 S87 T90	N88 P72 P80	P85	K86 S87 T90 T93
N87 P72 P77	H75	L91 R96 T93	P71 P72 P73	P98	P76
A23 H75 H86	H83	H86 K86 L84 L90 L91 P76 S87 T93	B74 N86 R95	R79	K86 P76 R95 R96 T93
F74 H75 H83	H84	K86	A79 K72 L71	R84	A79 R95 T90
H83 H84 L71	H86	K86 L91 T93	D76 L71 Y78	R85	
H83 P28 P85	K22	K74 K86 T90	K82 P17 R79	R95	K86 R96 T90 T93
A79 D76 X76	K72	A79 D75 K83 K92 K93 R84 S97 X76	H75 L71 P17 R79	R96	K86 L91 T90 T93
K22 K75 K76	K74	K75 K76 K78 K86 T93	F74 R85 U79	S77	K86
D74 D77 K74	K75	K76 K77 K86 K90 T90 T93	H83 P28 P85	S87	K86 L91 P76 S88
K74 K75 U79	K76	K77 K78 K83 K86 K92 T90 T93	B90 H83 S87	S88	K86 L91 S87 T93
F94 K75 K76	K77	K78 K83	S87 S88 U85	S91	K86 L91 T90 T93
F84 K76 K77	K78	K83 K86 K90	K72 P70 Y78	S97	T90
B73 K77 K78	K82	K77 K83 R95	H83 L71 N74	T71	K86 L91 N89 P76 R96 X76
K77 K78 K82	K83	K86	L88 P17 T85	T80	D93 K92 L91 T86 T90 T99
A23 B90 F94	K86	L91 T90 T93	B78 T71 T80	T85	K86 T86 T90 T93
P80 U85 U88	K87	K86 K92 T93	H75 T80 T85	T86	T90 T93 T99
K75 K78 U79	K90	K86 T90	A23 A90 F83	T89	K86 T90 T93
F74 K76 K90	K91	T90	D75 K22 T89	T90	K86 T93
K72 K87 T80	K92	T93	F74 L91 U79	T92	K86 L91 T90 T93
K72 P28 U79	K93	K94	K86 N70 T90	T93	K86 T90
K93 U79 Y78	K94	K86 L91	B78 T80 T86	T99	K86 L91
A79 D75 D76	L71	K86 L91 R84 R96	H83 N87 X77	U04	K86 L91
B78 H83 P28	L84	K86 L85 L89 L91 L95 T93	D74 T71 U79	U75	A79
F74 L84 N90	L85	K86 L89 L91 L95 P76 T93	U75 U77 U79	U76	K86 Y77
H83 L91 N85	L88	K86 L91 T90	U75 U76 U79	U77	A79 U76
L84 L85 P28	L89	K86 L90 L91	B73 K91 U75	U79	U75
F74 L89 P28	L90	K86 L91	K75 L85 U75	U85	K86 L91 T90
H75 L88 N70	L91	K86 T93	A90 B74 K87	U88	K86 T90 T93
A90 L84 L85	L95	K86 L91 T93	N86 R84 T71	X75	K86 P76 X76
F74 L85 X75	L98	K86 L91 P76 T93	K72 T71 X75	X76	K86
	N70	L91 T93	L71 X75 X76	X77	A79 K86
A23 D77 U79	N74	N88 T90	U75 U76 Y85	Y77	K86
H83 L85 P85	N85	K86 L91 P76 S87	B72 P77 U88	Y78	A79 D93 K94 P15 R96 T90
K87 K91 L71 N74	N86	K86 L91 P76 R96	L71 U76 U79	Y85	K86
K91 L71 P72	N87	P70 T90			

5.5.3 Supplementary figures

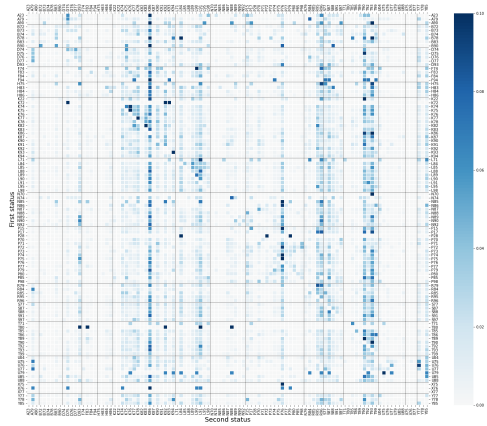


Figure 5.4: Subgroup analysis – Markov chain analysis -Male

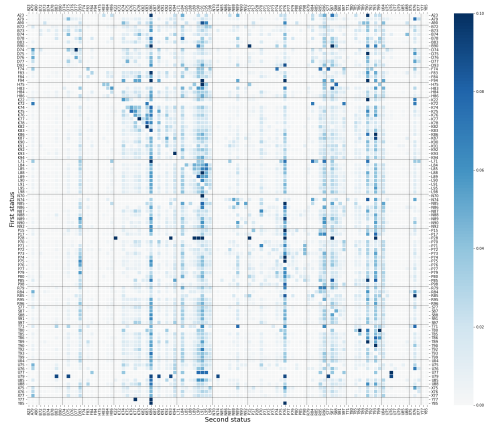


Figure 5.5: Subgroup analysis – Markov chain analysis -Female

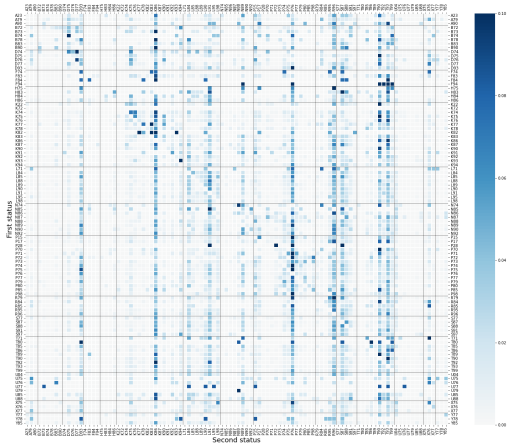


Figure 5.6: Subgroup analysis – Markov chain analysis - Age group 40-59 years old

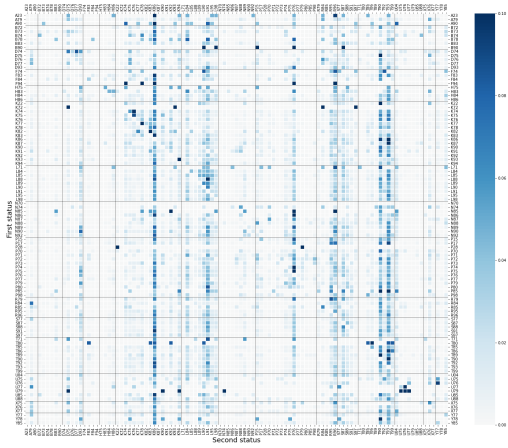


Figure 5.7: Subgroup analysis – Markov chain analysis - Age group 60-74 years old

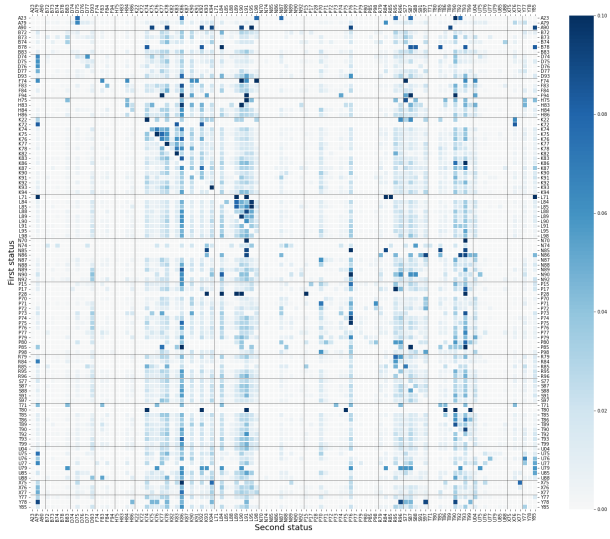
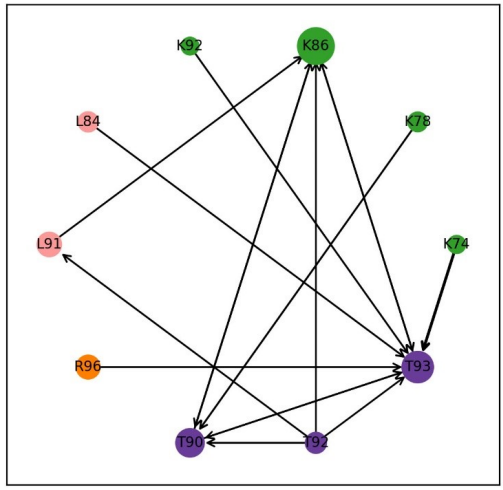
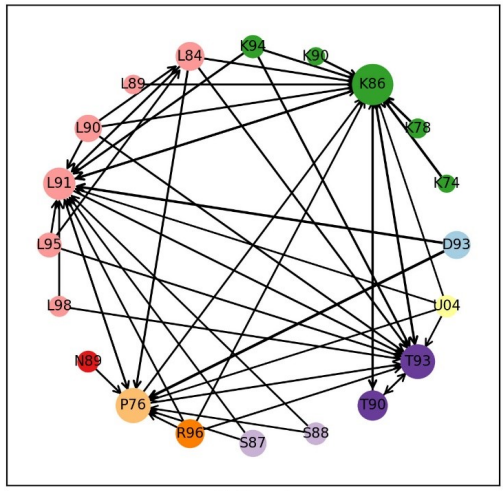


Figure 5.8: Subgroup analysis – Markov chain analysis - Age group above 74 years old



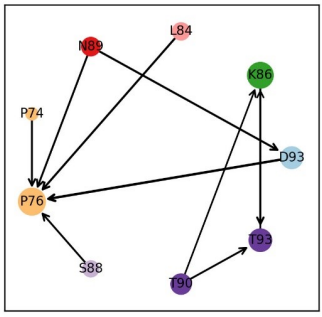
(a) Male



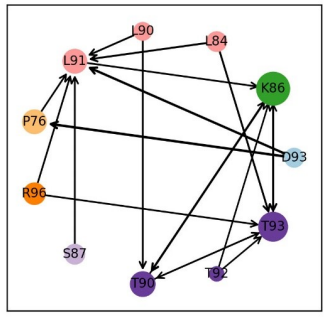
(b) Female

D93 Irritable bowel syndrome	L95 Osteoporosis
K74 Ischaemic heart disease w. angina	L98 Acquired deformity of limb
K78 Atrial fibrillation/flutter	N89 Migraine
K86 Hypertension uncomplicated	P76 Depressive disorder
K90 Stroke/cerebrovascular accident	R96 Asthma
K92 Atherosclerosis/PVD	S87 Dermatitis/atopic eczema
K94 Phlebitis/thrombophlebitis	S88 Dermatitis contact/allergic
L84 Back syndrome w/o radiating pain	T90 Diabetes non-insulin dependent
L89 Osteoarthrosis of hip	T92 Gout
L90 Osteoarthrosis of knee	T93 Lipid disorder
L91 Osteoarthrosis other	U04 Incontinence urine

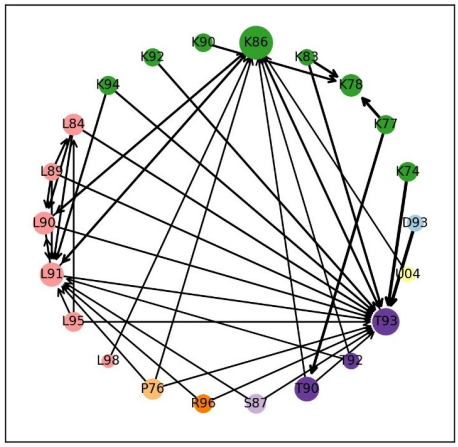
Figure 5.9: Subgroup analysis – WARM - Sex group



(a) 40-59 years old



(b) 60-74 years old



(c) 75 years old and over

D93 Irritable bowel syndrome	L90 Osteoarthrosis of knee
K74 Ischaemic heart disease w. angina	L91 Osteoarthrosis other
K77 Heart failure	L95 Osteoporosis
K78 Atrial fibrillation/flutter	L98 Acquired deformity of limb
K83 Heart valve disease NOS	P76 Depressive disorder
K86 Hypertension uncomplicated	R96 Asthma
K90 Stroke/cerebrovascular accident	S87 Dermatitis/atopic eczema
K92 Atherosclerosis/PVD	T90 Diabetes non-insulin dependent
K94 Phlebitis/thrombophlebitis	T92 Gout
L84 Back syndrome w/o radiating pain	T93 Lipid disorder
L89 Osteoarthrosis of hip	U04 Incontinence urine

Figure 5.10: Subgroup analysis – WARM - Age group

Chapter 6

System Architecture of a Clinical Decision Support System

This chapter is submitted to Journal of Medical Internet Research as:

Shi X, Nikolic G, Fischaber S, Black M, Rankin D, Epelde G, Beristain A, Alvarez R, Arrue M, Pita Costa J, Grobelnik M, Stopar L, Pajula J, Umer A, Poliwoda P,Wallace J, Carlin P, Paakkonen J, De Moor B. System Architecture of A European Platform for Health Policy Decision Making: MIDAS.

ABSTRACT

Background: Healthcare data is a rich yet underutilized resource due to its disconnected, heterogeneous nature. A means of connecting healthcare data and integrating it with additional open and social data in a secure

way can support the monumental challenge policy-makers face in safely accessing all relevant data to assist in managing the health and wellbeing of all.

Objective: The goal of this study was to develop a novel health data platform within the MIDAS (Meaningful Integration of Data Analytics and Services) project, that harnesses the potential of latent healthcare data in combination with open and social data to support evidence-based health policy decision-making in a privacy-preserving manner.

Methods: The MIDAS platform was developed in an iterative and collaborative way with close involvement of academia, industry, healthcare staff and policy-makers, to solve tasks including data storage, data harmonization, data analytics and visualizations, and open and social data analytics. The platform has been piloted and tested by health departments in four European countries, each focusing on different region-specific health challenges and related data sources.

Results: A novel health data platform solving the needs of Public Health decision-makers was successfully implemented within the four pilot regions connecting heterogeneous healthcare datasets and open datasets and turning large amounts of previously isolated data into actionable information allowing for evidence-based health policy-making and risk stratification through the application and visualization of advanced analytics.

Conclusions: The MIDAS platform delivers a secure, effective and integrated solution to deal with health data, providing support for health policy decision-making, planning of public health activities and the implementation of the Health in All Policies approach. The platform has proven transferable, sustainable and scalable across policies, data and regions.

Keywords: public health, decision support system, epidemiology, data visualization, machine learning

6.1 Introduction

We live in a data-rich society, which provides extensive opportunities for the development of big data and artificial intelligence technologies to provide new insights to enhance decision-making. Such technologies have particular importance in healthcare and health policy making. Despite the urgent need and opportunity, their use has not reached full potential in this field for various reasons, for example, healthcare data is typically heterogeneous and disconnected, existing in isolated silos, making meaningful analysis difficult. Privacy concerns create an additional barrier in exploiting the potential of healthcare data, preventing data sharing in a timely manner.

A means of connecting healthcare data and integrating it with additional open and social data in a secure way did not exist prior to the MIDAS platform release. Such a solution can support the monumental challenge policy-makers face in safely accessing all relevant data to assist in managing the health and wellbeing of all.

The MIDAS project set out to address this challenge and has developed a novel helath data platform that connects a range of heterogeneous health-related data with open and social data and applies advanced analytics techniques to provide a visual data-driven decision making tool that enhances healthcare policy making, whilst ensuring key aspects of ethics, security and privacy are adhered to [22]. MIDAS has successfully provided a tool that transforms the way health policies are developed, evaluated and implemented, meeting all of its specified objectives.

The platform has been piloted across four European regions: Finland, Ireland, Basque Country (Spain) and Northern Ireland (United Kingdom), addressing major health challenges in each region including mental health, diabetes and the ageing population, childhood obesity and social care for children, respectively.

This paper will present the system architecture of the MIDAS platform, which integrates data warehouse, data analytics, data visualization, and external applications for social media analysis, and enables rapid adjustments to new pilots.

6.2 System Architecture

The MIDAS platform was developed in an iterative and collaborative way with close involvement of policy-makers and experts who informed data exploration and analysis based on their expertise. The co-created platform solves the practical policy questions proposed by the policy-makers and provides the possibility of being applied to a wider range of topics in a generally automated process. Moreover, the MIDAS platform addresses the problem of how the data can be linked, harmonized, analyzed, and visualized in a multinational framework. The scope of the specification encompasses user-interface integration, authentication and authorization, data storage, data preparation, analytics backend, visualization, and connection with external resources.

6.2.1 Pilots

MIDAS was developed in the light of the needs of four very different pilot sites with different research topics and data sources, namely the Basque Country (Spain), Finland, Northern Ireland (United Kingdom), and Ireland. The research objectives for each pilot are listed in Table 6.1. The MIDAS project aimed to develop a platform that could deal with a wide range of topics in the international context using machine learning models. Therefore, each pilot had a unique research topic, and separate tailored dashboards were developed for them built upon a uniform architecture.

Table 6.1: Research topics for all pilots

Pilot	Research Objectives	Data Source
Basque	To understand what drives childhood obesity and the etiology of the childhood obesity	Controlled and open data
Finland	To understand mental health issues of young people with the support of visual analytics and analysis of available diverse datasets	Controlled and open data
Northern Ireland	To analyze the anonymized data extracted from children care system to provide new insights into a child' s journey through the care system	Controlled and open data
Republic of Ireland	To study the cohort of persons with diabetes and determine the best distribution for diabetes services	Controlled and open data

6.2.2 Platform Overview

As shown in Figure 6.1, the MIDAS Platform consists of a Policy Site Network and an External Network. The back-end analytics platform includes the Core Data Platform, tools for data harmonization (GYDRA), Analytics Backend, tools for data visualization (OpenVA), and three open and social data analytics and engagement tools in External Analytics Platforms. The policy-makers in a pilot site could adopt tools in the User Interface (UI), including the ISAACUS Metadata server, GYDRA, MIDAS Dashboard, News Media Dashboard, MEDLINE Dashboard, and Social Campaign Manager.

An overview of the Core Data Platform configured for MIDAS is given in Figure 6.2, including the core services of data storage and processing, user applications for interacting with these services, local user authorization and authentication, and data virtualization.

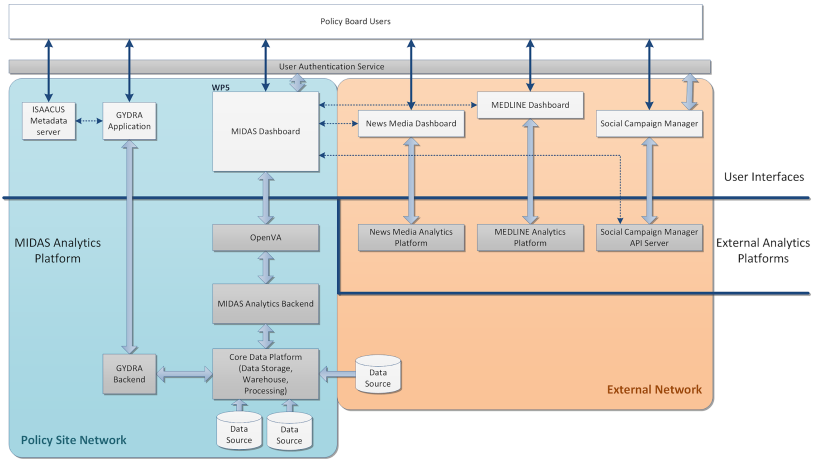


Figure 6.1: MIDAS Platform Overview. The platform consists of software hosted within a Policy Site Network for analyzing local data and applications hosted externally in External Network for analyzing open and social data as indicated by the colored boxes. The light-grey boxes indicate end user-facing web-applications, connected to back-end applications and analytics platforms shown in the dark-grey boxes. The white storages are the data sources, either internally from different pilots or externally from open and social data. The arrows between different softwares indicate they have direct interactions.

6.2.3 Data Preparation and Harmonization

The data preparation and harmonization task aimed to develop appropriate pre-processing modules for preparing the raw data to ensure that they were compatible with the agreed data representations and could be used for analysis, including for instance: data cleansing, normalization, transformation, joining, and missing value imputation. The GYDRA software (renamed from TAQIH) was developed and applied for data preprocessing and transformation [150, 54]. The GYDRA is a customizable tool for facilitating the data wrangling process through interactive and visual tools, taking advantage of machine learning algorithms. The aim is to

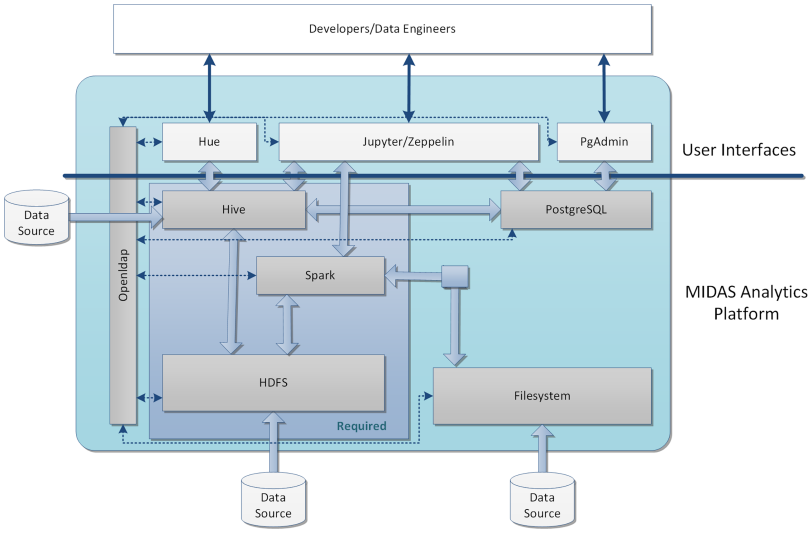


Figure 6.2: An overview of the system configured for MIDAS. The core data platform for the MIDAS stack was based on HDFS, Hive, and Spark. The data can be imported into the system through Filesystem, HDFS, or externally to Hive. HDFS was applied to store files and raw data and Hive was employed as a data warehouse for the structured data after processing. External data assets were also virtualized through the Hive interface and they could be accessed by the MIDAS tools similarly to locally loaded data assets and used within GYDRA. The UI of the analytic platform includes Jupyter Notebook with Python and PySpark for developing and testing the underlying analytics models before being implemented within the MIDAS Analytics Backend. For managing and querying the databases in Hive and PostgreSql, an open sourced interactive editor Hue was used, and Zeppelin provided support for running Spark applications. User access was managed by a local LDAP server, which provided role-based access to the user applications and underlying data sources.

simplify the tedious and time consuming part of data analysis, allowing non-technical users to transform raw data into information ready for analysis.

The GYDRA provides web interfaces to understand the content, structure and distribution of the dataset through an easy-to-use tab-based navigation approach following common data assessment and preparation steps. Figure 6.3 presents screenshots for two representative sections for general statistics and missing values, respectively. Moreover, on each tab or section of the application, a visual transformation pipeline allows the users to add a dataset transformation action after knowing the dataset content.

As a python-centered solution, with an easy-to-use interactive UI, the GYDRA uses Celery for asynchronous distributed data-processing suitable for handling big HDFS datasets that do not fit into system memory. Additionally, through the web-based GYDRA tool, a data synchronization function allows the data owners and policy-makers to efficiently deploy prepared datasets to the analytics platform. The synchronization logic aligns the GYDRA metadata tool with the ISAACUS metadata server (available on https://github.com/THLfi/thldtkk) and updates the data warehouses further through the GYDRA backend depicted in Fig. 1. The raw data of each health policy area was prepared and processed using the GYDRA tool thus making the data ready for the MIDAS Analytics Backend.

6.2.4 Data Analytics

The MIDAS Analytics Backend provides the back-end analytics and simulation results required for the MIDAS dashboard. Apart from being a middle layer linking the data preparation and the data visualization, it supervises the user in selecting the correct data tables and data variables for chosen analytics and visualization scenarios.

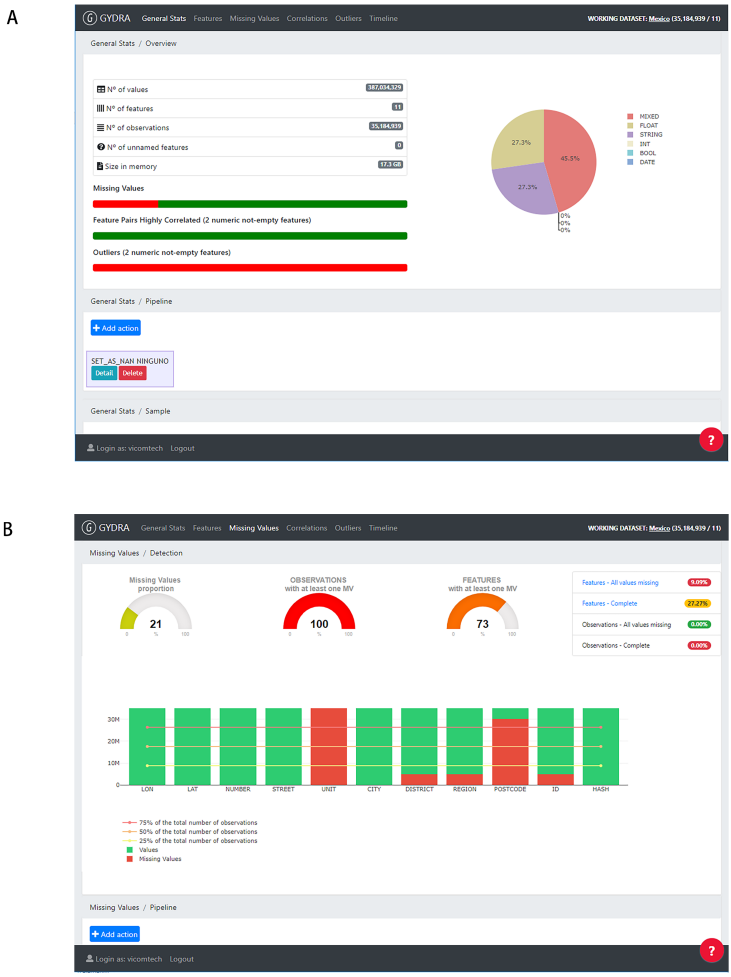


Figure 6.3: GYDRA data preparation tool UI, provided through an easy-to-use tab-based navigation approach following a common data assessment and preparation steps (i.e. General Stats, Features, Missing Values, Correlation and Outliers analysis tasks). Screenshots for two representative sections are included: (A) General stats - On the left side, general statistics on features and observations are provided, on the right side the variables type distribution is shown on a pie chart. On the lower part a transformations pipeline is included to add dataset transformations as their need is identified. (B) Missing values - On the top left area, missing value proportion is depicted for values, features and observations, on the right side indicators for complete and completely empty features and observations is provided. On the lower part, each feature is analysed separately on bar charts representing their missing value percentage.

The communication between the analytics and visualization layers was managed through a REST API server developed with the Flask microframework for Python. The Analytics APIs were developed to support generic exploratory data analysis (EDA), uniform across all pilot sites, as well as more specialized cross-filter dashboards and health policy simulators specific to each pilot-site.

The EDA was uniform for all pilots, providing eight types of basic visualizations for the selected variables from the harmonized data, i.e. scatter plot, heatmap, histogram, bar chart, pie chart, bubble plot and choropleth map. The cross-filter analytics for each pilot platform (Figure 6.4), which are interactive visualization tools, had the same basic principle to update their content when the user selects different values on the displayed graphs. The associated visualizations were flexible for different pilots, with the layout and categorical variables proposed by policy-makers, including components such as line chart, bar chart, and tables.

Different machine learning methods were applied for each pilot to solve their unique research questions (Table 6.2). Because of data protection regulations, the data-related results cannot be shown. As the main focus of the paper is on system architecture, the detailed results are not discussed and shown in this paper.

The private MIDAS GitHub repository contains branches of each pilot, consisting of API endpoints for generic EDA, cross-filter, and pilot-specific analytics. Different types of cross-filter and pilot-specific analytics were deployed on each of the pilots in an iterative process. Each deployment iteration consisted of getting feedback from policy-makers on required analytics with evaluation of results through various technical, integration and face-to-face meetings, making it possible to meet the real needs of the policy-makers.



Figure 6.4: Cross-filter of the Finnish Pilot. Users can select the county or region in the map on the right and all the charts will update automatically. Gender, Region, and Age Group are the categorical variables shown in the lower panel next to the map, and users could select subgroups either by pressing the buttons on the panel or selecting the subgroups in the line chart or bar chart. The Finnish cross-filter consists of a matrix heatmap, a bar chart, a line chart and the regional map of Finland. The cross-filter of other pilots have different components.

6.2.5 Data visualization

Data visualization was deployed utilizing a three-tier architecture in the MIDAS Dashboard, including the MS Azure AD B2C authentication service [12], the OpenVA middleware framework [2], and the dedicated MIDAS UI single page application (SPA), which provides decision-making support for policy-makers with data-driven analytics from the internal and external resources.

A Single-Sign-On service was implemented between the MIDAS Dashboard and external resources through the common authentication service, mentioned above. The OpenVA framework handles the connectivity of shared SPA to local resources and dedicated external components. The MIDAS UI SPA is shared by all instances from a centralized web server and

Table 6.2: Pilot-specific analytics

Pilot	Method	Purposes
Basque	Random Forest-s/LASSO	To identify the risk factors of childhood obesity
Finland	Lexis diagram analysis	To aggregate, summarize and visualize the selected risk factors in a secure way to protect the privacy of patients
Finland	Descriptive analysis	To intuitively evaluate the health, social, and education status of the inhabitants in regional level on a yearly basis by using open data
Northern Ireland	Markov chain	To track patterns of behaviour over time and to give better visualization to present intuitively how children move in and out of different types of care by estimating the probability of the transition between different types of care
Northern Ireland	LSTM Network	To predict future status of children to improve the protection for children from the policy level
Republic of Ireland	ARIMA	To forecast the consumption of diabetic drugs

it connects to the local OpenVA instance in line with the account details of the current user. The external resources include the Social Campaign Manager, MEDLINE Publication search and News Media search dashboard.

Through the MIDAS UI, users can generate a dashboard and interact with a widget wizard to generate the specific visualization widgets that can help them with policy decision support. Furthermore, additional pilot-specific dashboards and analytics tools were developed for each pilot, supporting each user in exploring and understanding their main research question. The MIDAS UI (Figure 6.5(A)) shows the visualized analytics results for selected datasets, together with the reporting tool illustrated in Figure 6.5(B) to allow users to generate a PDF report.

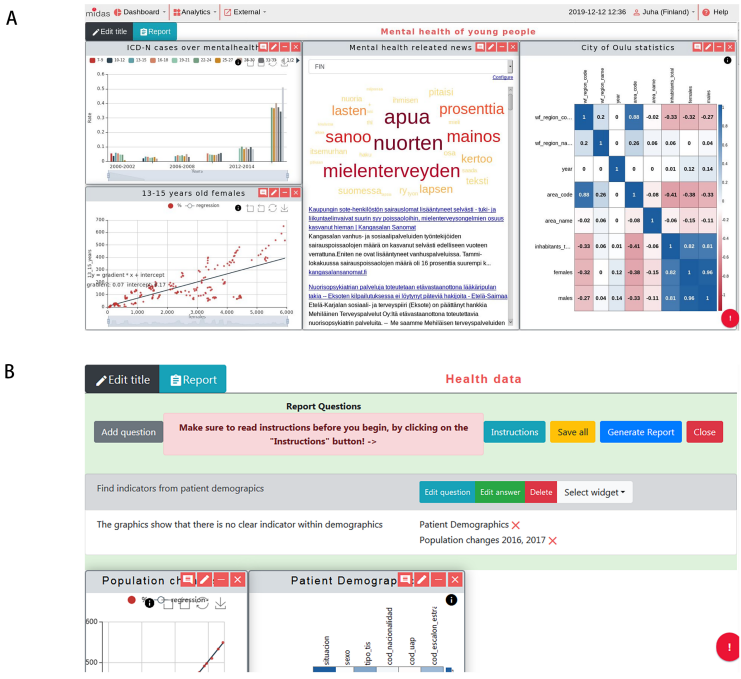


Figure 6.5: MIDAS UI screenshots. (A) Common view of generated dashboard. On top of the view are the menus to manage dashboards, add analytics and use the external resources. The reporting tool is located in the middle (hidden in figure) and the rest of the view is the open space for widgets. Users can freely resize and organize widgets in this space. The analytic results come from the MIDAS Analytics Backend, together with the widgets developed externally. (B) Reporting tool open with a single research question and an answer with two associated widgets. This tool is used to generate a PDF file, defining the research questions and answers and attaching the most suitable figures describing them from the available widgets.

6.2.6 Open and Social Data

Social Media Analysis

Social media is considered as an important source of information for policy making, to help better understand motivations and determine public acceptance of implemented policies. The social media analysis provides insights into the public' s perception and sentiments towards health policies by having members of the public engage with the created chatbots through a series of questions about a specific health policy. The user responses are analyzed in real time as they are entered into the system using IBM Watson Natural Language Understanding (NLU) APIs, and the sentiment and emotional analysis are then displayed on the Social Campaign Manager dashboard or widget on the MIDAS UI dashboard (Figure 6.6(A)).

The Social Campaign Manager was hosted as a microservice on the IBM Cloud platform. Twitter was used to interact with the public and the IBM Watson Assistant and Watson NLU services were used for the chatbot. The Social Campaign Manager was a separate web application for creating, running, and managing the individual campaigns. These provided the intelligence and dialog capability to interact with the user as well as performing the analysis of the conversations. The Social Campaign Manager API Server was the core application connecting these various services, providing data to the MIDAS Dashboard and the Social Campaign Manager web application. The authentication used within MIDAS platform layers is OAuth 2.0, a common industry-standard protocol for authorization.

MEDLINE Analytics

The MEDLINE dashboard accessible through the MIDAS platform was developed to provide dedicated text-mining tools and visualizations to enable users to extract meaningful information from the MEDLINE dataset

[65, 165]. The MEDLINE dataset was indexed using the ElasticSearch, and visualized through an open source tool Kibana [89]. The purpose of the dashboard was to provide users with tools to explore the insights of published biomedical research, in an intuitive manner. The main advantage is the dynamic article prioritization (ranking). The user enters a few keywords in the search box and results are shown (Figure 6.6(C)). This visual interactive widget helps surface information that one is looking for by re-ranking the top 10 articles, letting the users interact with the index of the results, getting them closer to the scientific information that they are looking for.

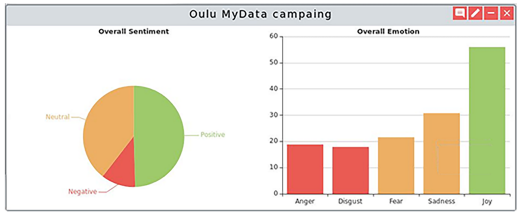
Each topic dashboard was developed through extensive interactions with the pilot sites, improving the understanding of how the tools could be used to address specific use cases. The MEDLINE knowledge was also served directly at the MIDAS platform by a widget that also allowed for Lucene queries and for the user to interact with a pointer over a tag cloud of related topics to alter the order of scientific articles provided as result of the query.

News Media Analysis

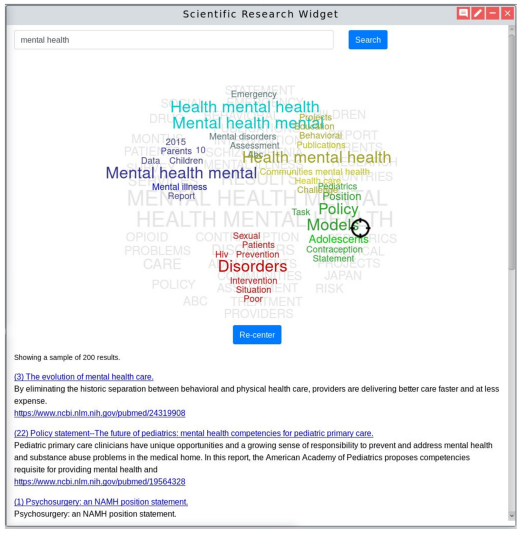
MIDAS provides users with tools to monitor specific health topics in the worldwide and local news. The news media analysis tool is available through the platform (Figure 6.6(B)), enabling the monitoring of worldwide news outlets and the enriching of these news articles with data from the MEDLINE knowledge base [37]. Each pilot region in the MIDAS project has its own live news source which can be accessed via the dedicated news data exploration dashboard served by the Event Registry news engine and through a widget within the MIDAS Dashboard UI [97]. In addition to setting up the pilot-specific data streams, the underlying data sources for Event Registry were improved to better support Finnish and Basque language news coverage, adding to the 60+ languages available. In addition to the news media tools, a MeSH Classifier tool was developed which enables classification of news

articles (and any text snippets) with MeSH terms. The system is available through a web portal and a REST API, and includes a NodeJS wrapper for direct inclusion into other systems [38].

A


Oulu MyData campaing
Overall Sentiment
Neutral
Positive
Negative
Overall Emotion
60
50
40
30
20
10
0
Anger
Disgust
Fear
Sadness
Joy

B


Scientific Research Widget
mental health
Search
Emergency
Health mental health
Mental health mental
Mental disorders
Assessment
Behavioral
Publications
2015
Parents 10
Data Children
Health mental health
Mental health mental
Communities mental health
Mental illness
Report
Pediatrics
Position
Task
Policy
Models
Sexual
Patients
Hiv Prevention
Adolescents
Contraception
Statement
Disorders
Intervention
Situation
Poor
Re-center
Showing a sample of 200 results.
(3) The evolution of mental health care.
By eliminating the historic separation between behavioral and physical health care, providers are delivering better care faster and at less expense.
https://www.ncbi.nlm.nih.gov/pubmed/24319908
(22) Policy statement--The future of pediatrics: mental health competencies for pediatric primary care.
Pediatric primary care clinicians have unique opportunities and a growing sense of responsibility to prevent and address mental health and substance abuse problems in the medical home. In this report, the American Academy of Pediatrics proposes competencies requisite for providing mental health and
https://www.ncbi.nlm.nih.gov/pubmed/19564328
(1) Psychosurgery: an NAMH position statement.
Psychosurgery: an NAMH position statement.

C

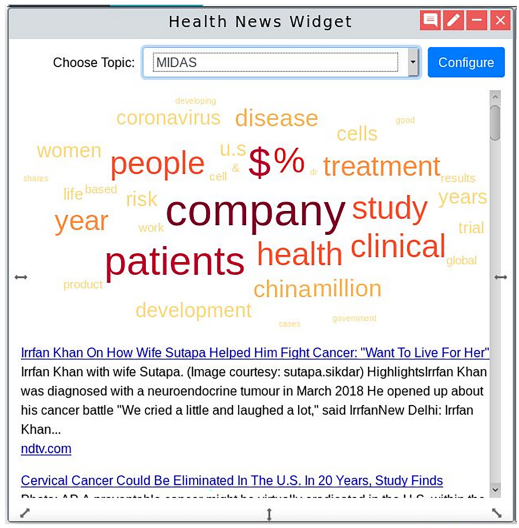

Health News Widget
Choose Topic:
MIDAS
Configure
coronavirus
disease
cells
women
people
$%
treatment
life
risk
company
study
years
year
trial
patients
health
clinical
chinamillion
development
Irrfan Khan On How Wife Sutapa Helped Him Fight Cancer: "Want To Live For Her"
Irrfan Khan with wife Sutapa. (Image courtesy: sutapa.sikdar) HighlightsIrrfan Khan was diagnosed with a neuroendocrine tumour in March 2018 He opened up about his cancer battle "We cried a little and laughed a lot," said IrrfanNew Delhi: Irrfan Khan...
ndtv.com
Cervical Cancer Could Be Eliminated In The U.S. In 20 Years, Study Finds

Figure 6.6 *(previous page)*: Open and Social Analytics(A) Social Campaign Manager shows a high-level overview of the (i) sentiment and (ii) emotional analysis of the policy being studied by a Twitter social media campaign. The sentiment and emotions found in responses to a public online survey reaching out to the public to gather their voice on a specific health policy being considered on the dashboard. Clicking into the dashboard provides further insight including responses to particular questions and results processed using Natural Language Processing techniques showing the most common topics of conversation mentioned in the responses and the sentiment in which they were made. (B) News custom widget that consists of: (i) a word cloud that represents the main topics of the listed news, enabling a global perspective of the key topics before further activity; (ii) a list of news titles and first lines that are linked to the original news source; and (iii) the search choices where news are based on, defined by the filter and search options at the "Media Monitoring" menu of the external news dashboard. (C) MEDLINE custom widget that includes: (i) a list of the top ten MEDLINE articles with the first part of the abstract serving as a short description; (ii) a tag-cloud representing clusters of topics extracted from the MEDLINE articles including the searched keywords; and (iii) a target-shaped pointer that the user can move through the tag-cloud and by that, change the ranking of the listed articles.

6.3 Results

The developed MIDAS platform consists of several dashboards, including Exploratory Data Analysis (EDA) (shown in Figure 6.5), cross-filter dashboard (Figure 6.4), pilot-specific analytics dashboard, and social media dashboards (Figure 6.6).

The platform development followed an agile approach where iterations of the platform were evaluated at periodic milestones and feedback acquired was incorporated to produce the final product [34, 33]. We conducted 3 rounds of user experience testing to help improve the intermediate prototypes. The participants included data analysts and policy makers, a more detailed demographic statistics can be seen in Table 6.3. We guided the participants

to finish a list of tasks and collected their feedback and suggestions for further improvements. The user experience testing helped successfully identify the potential problems, and improvements were achieved after incorporating user feedback.

Table 6.3: The background information of the participants of user experience testing

	Version 1.0	Version 2.0	Version 2.1	Version 3.0
N	12	5	7	3
Age	-	47.2	51.9	52.0
Gender				
-Female	7	2	5	1
-Male	5	3	2	2
Occupation				
-Data Analyst	7	5	4	2
-Policy Maker	5	0	3	1

The final versions of the pilot platforms were evaluated by policy makers from all pilots based on the Key Performance Indicators (KPI) (Table 6.4). The second column is the demands proposed by policy makers, and the third column is the corresponding function on the MIDAS platform. All KPIs were successfully achieved and the platform has received positive feedback from stakeholders on its capacity to integrate and analyze previously fractured heterogeneous data. Furthermore, the ability to produce new knowledge and results that are actionable by health policy-makers was demonstrated. The custom-tailored analytics solved the practical questions for the health policy-makers and gave them insights for possible future interventions. The platform can be easily manipulated by users without technical background by following the User Guide [144].

6.4 Discussion

6.4.1 Principal Results

The core user groups of these tools are mainly business users, dashboard users and in-house analytics teams. In contrast, the MIDAS platform was co-created by academia, industry, and crucially, healthcare staff, health policy-makers, patients and citizens thus ensuring the solution' s design and development has been user-led. With this user-centered approach, the MIDAS platform guides its users through all steps of the data analytics pipelines. Besides, data blending is restricted according to the prior knowledge of the original data. These restrictions assist the user in selecting only suitable variables for a chosen visualization, thus producing meaningful analytics and visualization results.

Because of the flexibility of the open data tools, they can be quickly adjusted to study the most urgent topics, as a result, MIDAS recently presented a fast response to the COVID-19 global initiative [135]. This impactful public health event was addressed through the worldwide news, offering the customized news streams through the MIDAS news widget, to help the pilot site use cases to better track news and relate it to their own priorities.

In order to maximize the sustainability of the MIDAS platform beyond the lifetime of the project, we explored a range of mechanisms for coordinating further development and marketing activities among the project contributors post-project. After detailed partner and stakeholder engagement we determined that the establishment of a MIDAS Open Source Foundation would be the most suitable approach. New regions, cities, and organizations from Scotland, France, and Spain have confirmed their interest, with more public sector policy departments noticing the platform capability of addressing similar problems in their area in future.

6.4.2 Comparison with Prior Work

The MIDAS platform tries to maintain the privacy of each stakeholder by keeping their sensitive health data in-house. Other commercial tools like Tableau, PowerBI or QlikView often require a connection to external services, while all layers of the MIDAS platform are hosted inside the stakeholder' s trusted zone. Moreover, they are general purpose solutions that do not consider the specific challenges of public health data, nor the user stories of the target MIDAS audiences. Therefore, each layer of the MIDAS platform supports a secure data analytics pipeline and minimizes data-leakage. Additionally, the learning curve of some commercial tools can be steep, requiring specialized training. In terms of advanced analytics capabilities, Tableau provides some advanced analytics features but with external integration, PowerBI has core competency and integration, while QlikView does not offer any advanced analytics features.

6.4.3 Conclusion

This study has demonstrated the value of a secure, effective and integrated solution that deals with health data to harness the potential of underutilized healthcare data and provide support for health policy decision making. The MIDAS platform was successfully implemented within the four pilot regions and has received positive feedback from stakeholders on its capacity to turn large amounts of previously isolated data into actionable information to inform health policy making and risk stratification through the applications and visualizations of advanced analytics. By delivering the MIDAS platform as an innovative and state-of-the-art solution, we have successfully provided a tool that can potentially transform the way health policies are developed, evaluated and implemented, which will ultimately enable impactful improvements in public health and the quality of life amongst European citizens and beyond. Besides, the platform has successfully

demonstrated that it is transferable, sustainable and scalable across policies, data and regions.

Table 6.4: Research topics for all pilots

Pilot	Updated KPI	MIDAS Platform data and components to meet the KPIs
Basque	KPI 1: To provide and clearly identify both crude and adjusted rates. KPI 2: To provide information at different granularity levels, adequate for each stakeholder. KPI 3: To cover time, place, and person analysis for epidemiological studies	KPI 1: These were calculated by cross-filtering tool as new diabetes cases adjusted rates per age, sex, region, area and registration year. KPI 2: Basque cross-filtering tool user interface was developed to answer this. OSI (trust) and UAP (primary care unit) level. KPI 3: Also this is covered by Basque cross-filtering tool with gender and age filters (person dimension) over different years (time dimension) and for two different granularity level health area distributions (place dimension)
Finland	National level: KPI 1: Support understanding and managing the challenges of mental health issues with young people. KPI 2: Enable analysis of the available datasets to identify the leverage points in effective preventative measures. KPI 3: Enable the use of visual analytics to assess the impact of different factors and indicators for policy decisions. Regional level (City of Oulu): KPI 4: Enable city data and open data evaluations from various sources in social and healthcare units in the stakeholder organizations to support population and area comparisons for preventive policy making.	National level: KPI 1: Main contribution is given by the cross-filtering tool for Lexis analysis results. KPI 2: The system dynamic model developed for the national level research question and procedure to create the model KPI 3: Lexis analytics related single Lexis result visualizations on Dashboard and simulator interface of system dynamic model Regional level (City of Oulu): KPI 4: City of Oulu level open data was collected to answer this question. The data is visualized with cross-filtering tool and single data visualizations options on MIDAS Dashboard. The young people's personal opinions from the Northern Finnish regions were covered by the NFBC dataset.
Northern Ireland	KPI 1: Analysis of the available datasets to identify effective preventative measures KPI 2: A longitudinal analysis and track a cohort of Looked After Children as they move in and out of care, use a variety of health services, and look at KPI 1: Analysis of the available datasets to identify effective preventative measures KPI 2: A longitudinal analysis and track a cohort of Looked After Children as they move in and out of care, use a variety of health services, and look at patterns of behaviors and changes over time. KPI 3: The use of visual analytics to assess the impact of changing variables and indicators on policy decisions KPI 4: Harmonization and integration of multiple data sources	KPI 1: Done within data integration work. Data analytics include LSTM simulator that allows policymakers to identify children at high risk. Analytics for Department of Health (DoH) allows exploring different subgroups of children per selected categories. KPI 2: Answered by pilot-specific analytics and related visualizations on MIDAS Dashboard: Markov Chain analysis, LSTM simulator and DoH analysis KPI 3: Visualization of KPI2 results KPI 4: Done as a part of data integration work
Republic of Ireland	KPI 1: The provision of an adoptable intelligent analytics platform for stakeholder use KPI 2: Identification of the cohort of persons with diabetes in the Republic of Ireland KPI 3: Determining of information that can be shared to facilitate better use of resources and services for people with diabetes KPI 4: Determining the best fit for diabetes services and their locations based on population based geographical needs KPI 5: Identification of improved outcomes if certain 'data insight' based policies are implemented	KPI 1: MIDAS Platform integration on pilot environment KPI 2: Dedicated tables and variables to identify these cohorts were created for the database KPI 3: The cross-filtering tool was developed to answer the (diabetes) drug usage questions KPI 4: Partially answered by the cross-filter tool geographical representation of drug usage KPI 5: This KPI was found not possible to answer as related data is not available for the project

Chapter 7

Conclusion

In this thesis, we have presented how machine learning methods can provide supports for clinical decision making in different clinical topics. In this chapter, we will summarize the contributions that have been achieved, and the follow-up work that can be done in the future.

7.1 Contributions

Our thesis covers the full process of clinical decision making, including medical data cleaning, risk factor selection, disease prediction, disease sequence analysis, and the system architecture of a clinical decision support system.

We proposed an automated data cleaning method for clinical variables. One of the main contributions of this method is that it allows non-technical people to explore real-world data without needing proficient data processing techniques. Despite the high automation is at the cost of slight decline in completeness, it is still reliable to get results of general trends and

associations. Another contribution of this method is it largely improved time efficiency with a slight loss in accuracy. It took around 5 mintues to clean around 50 variables by using our method, when it can take a proficient data scientist about 40 hours on average to clean these variables manually.

In terms of feature selection method, we were one of the first to apply the heterogeneous-distributed implementation for ensemble feature selection. Although some studies applied distributed implementation for individual methods [142], few papers tried it for ensemble feature selection, except some papers using homogeneous-distributed ensemble methods [155, 25, 113]. Another contribution is the voting strategy, which can also be regarded as the way to combine the outputs of ensemble feature selection. In order to achieve a feature set with better clinical relevance, the commonly used way was to combine the results by having the intersection or union of the features from different selectors [101], and some studies incorporate feature rankings to assign weights [171, 6]. Our voting strategy has both feature rankings and model performance, making it possible to select a more reasonable set. Moreover, the interpretability of machine learning methods are gaining more attention in recent years, but no discussion was found concerning the interpretability of algorithm-identified risk factors. We explored the possibility of selecting the more clinically meaningful feature set, and summarized the dimensions of interpretability for algorithm-identified features to better evaluate the results.

As for the prognostic predictive models, the imbalanced data structure caused by low prevalence of some diseases were rarely discussed in previous studies. Survival models or logistic regression were applied in the most studies, with AUC or C statistics as evaluation measures. However, as discussed in Chapter 4, these methods and evaluation measures are unable to give the correct prediction as expected. We suggested applying resampling method before the use of logistic regression to solve the problem of imbalanced data structure. It was not difficult to introduce the resampling

methods into the disease prediction, as it was more like a data preprocessing procedure, instead of adding complexity to the model. But the simple action solved the problem of imbalanced data structure, making the logistic regression able to detect the real ESRD patients. Apart from the resampling method, we also pointed out that evaluation measures could not reflect the real performance, especially when False Negative and False Positive should not be treated equally. After comparing the commonly used evaluation measures, such as accuracy, AUC, C statistics, Recall, Precision, we concluded that F_β score was the most suitable evaluation measure, with the parameter β representing the level of penalty of False Negative. The studied disease in the use case of prognostic models was chronic kidney disease, however, the same problem of imbalanced data structure can be commonly observed for most chronic diseases. Both the resampling techniques and the evaluation measure can be widely used for the studies on other diseases in the future.

It is a much more complicated situation where the patients have multiple diseases at the same time, which is called as multimorbidity. Despite the complexity and urgent need, there were not sufficient methods to completely explore this topic. To our knowledge, there were no studies about multimorbidity taking sequence into consideration before. In our study, the Markov chain analysis and WARM were used, making it one of the first to investigate multimorbidity considering the sequence of evolution. It is worth noticing that both Markov chains and WARM have rarely been applied in epidemiologic studies, although they are very useful techniques to explore sequential rules. Two papers have used ARM to summarize multimorbidity patterns [201, 71], but no studies have applied WARM so far. Besides, we presented a comprehensive view of the associations among 103 chronic conditions, covering chronic conditions from all disease groups, which might shed light to some potential relations that were not widely discussed before.

Finally, the system architecture of a platform was shown in Chapter 6, which could provide supports for health policy making. This chapter presents how a platform was developed from an implementation perspective and gives some snapshots of the user interface. It is possible to add additional components or applications on this basis in the future.

The thesis has a focus on improving interpretability of all the steps of the proposed CDSS, with interpretability defined as a broader concept for the target population. The automated data cleaning method (Chapter 2) and the predictive model (Chapter 4) had a higher level of clinical relevance, as the methods took the clinical practice into consideration when developing and selecting the methods. The feature selection method (Chapter 3) defined the dimensions of interpretability in the context of health policy decision making and evaluated the results correspondingly. The sequence analysis in Chapter 5 visualized the results using interpretable figures, and the results from the rule-based models are easier for the readers to understand compared to the clustering used in previous studies.

7.2 Future work

The whole process of clinical diagnostics is discussed in this thesis, with each chapter focusing on one specific step, i.e. data cleaning, feature selection, disease prediction, sequence analysis, and system architecture. The methods proposed in each chapter have their own weakness limited by time, resources, and the current related studies. It is expected to further improve these methods in future research.

One limitation of the automated data cleaning method proposed in this thesis is that data longitudinality and patient variability are not considered, making it unable to check and improve data consistency. By considering the longitudinal data from the same patients, the correctness of some variables

may be further improved. In order to achieve this, the clinical knowledge database needs to be expanded to include more information. The current study is the initial exploration of using clinical knowledge database to achieve automated data cleaning for EHR data. Therefore only the most basic cleaning procedures were included. In future research, we can add more clinical knowledge to enable more cleaning procedures, and even more complicated "IF-THEN" rules to help improve consistency. In terms of data longitudinality, it is reliable to add longitudinal information following some pre-defined rules. Such longitudinal rules can also be added into clinical knowledge database. Defining rules to detect abnormal values in the longitudinal data requires extensive discussion with data and clinical experts, which is expected to be done in future studies.

In Chapter 3, one ensemble feature selection method, BFSMR, was proposed. However, the predictive performance of the five models in the section of results was not very satisfactory due to the sparse structure, missing values, and the small number of selected features. The model performance can be further improved if more features are kept, or by using methods that are specifically for sparse data. This is not the focus of this thesis, but can be explored in future studies to show the flexibility of BFSMR. Chapter 3 also summarized the dimensions of interpretability for algorithm-identified risk factors, but the evaluation can only be done manually and qualitatively. It is necessary to continue looking for or creating new evaluation criteria which can be used to evaluate the interpretability quantitatively.

The validity of the prognostic model in Chapter 4 may be questioned because of the high misclassification rate of negative cases. However, in clinical practice, it might be more important to accurately detect ESRD patients than asking more patients to take good care of themselves when they were at low risk, because the medical care to prevent ESRD does not require additional medication and only some home tests are necessary. The misclassification rate of negative cases can be improved in future work if

more machine learning classification models are tested. The current study uses logistic regression after the resampling method, mainly because logistic regression has a similar structure to survival model. The goal of the study was to make it easier for the readers with clinical backgrounds to accept the idea of imbalanced data structure and understand the results from survival models were not as valid as they expected. Thus, other machine learning models, such as SVM and deep learning methods, were not tested and compared. But this can be done in the future work. Moreover, the subsequent analysis can be done for the identified CKD patients after tracking them for a period. The task of ESRD prediction can be broken down into 2 parts, the first part is to accurately identify positive cases using the method in Chapter 4, then further distinguish positive cases and negative cases among the "predicted negative" cases in step 1. The data used in this study was extracted from the primary care data which was collected without a focus on a specific disease. Once the high-risk CKD patients are identified, with more frequent and maybe more specific examinations, it is possible to further predict the true positive cases much more accurately from all identified CKD patients.

In Chapter 5, the findings of multimorbidity using our methods are in agreement with previous clinical findings, meaning that these approaches were valid and could be used on larger datasets in future studies, such as National Healthcare Systems or large private insurers. Because of the intensive information stored in the results of the Markov chain and the WARM, it is worth examining the results more carefully to discover potential new clinically relevant relations that have not been widely discussed before. One limitation of the study is the lack of a time concept in the analysis. Although the sequence of occurrence is considered, it is not possible to incorporate exact interval time between diagnoses. Finally, the results of the Markov chain and WARM could suggest strong associations, but causal relationships remain unclear and further studies are needed to confirm causality. Thus, more studies using new methods are still expected in the

future work to overcome these limitations.

Bibliography

[1] Meaningful integration of data, analytics and services. https://cordis.europa.eu/project/id/727721. Accessed April 10, 2021. pages 18

[2] Openva. https://github.com/pekka-siltanen/vttopenva. pages 141

[3] dataqualityr. https://rdrr.io/cran/dataQualityR/man/dataQualityR-package.html. Accessed July 09, 2021. pages 42

[4] Imbalanced data. https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data. Accessed December 21st, 2020. pages 81

[5] *Digital Design and Computer Architecture*. Elsevier, 2007. doi: 10.1016/b978-0-12-370497-9.x5000-8. URL https://doi.org/10.1016/b978-0-12-370497-9.x5000-8. pages 74

[6] S. Aerts, D. Lambrechts, S. Maity, P. V. Loo, B. Coessens, F. D. Smet, L.-C. Tranchevent, B. D. Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544, May 2006. doi: 10.1038/nbt1203. URL https://doi.org/10.1038/nbt1203. pages 13, 70, 77, 156

[7] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993. pages 101

[8] M. A. Ahmad, C. Eckert, and A. Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, Aug. 2018. doi: 10.1145/3233547.3233667. URL `https://doi.org/10.1145/3233547.3233667`. pages 5

[9] S. Y. Ahn, D. K. Kim, J. H. Park, S. J. Shin, S. H. Lee, B. S. Choi, C. S. Lim, A. Lee, H. Jung, and H. J. Chin. Long-term effects of intensive low-salt diet education on deterioration of glomerular filtration rate among non-diabetic hypertensive patients with chronic kidney disease. *Kidney and Blood Pressure Research*, 44(5):1101–1114, 2019. doi: 10.1159/000502354. URL `https://doi.org/10.1159/000502354`. pages 92

[10] E. P. Ambinder. Electronic health records. *Journal of Oncology Practice*, 1(2):57–63, Jul 2005. pages 3

[11] M. Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015. pages 116

[12] M. Azure. Azure active directory b2c. `https://azure.microsoft.com/en-us/services/active-directory-b2c/`. pages 141

[13] F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, and D. Khalili. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology*, 71:76–85, Mar. 2016. doi: 10.1016/j.jclinepi.2015.10.002. URL `https://doi.org/10.1016/j.jclinepi.2015.10.002`. pages 51

[14] J. A. W. Baidal, L. M. Locks, E. R. Cheng, T. L. Blake-Lamb, M. E. Perkins, and E. M. Taveras. Risk factors for childhood obesity in the first 1, 000 days. *American Journal of Preventive Medicine*, 50(6):761–779, June 2016. doi: 10.1016/j.amepre.2015.11.012. URL https://doi.org/10.1016/j.amepre.2015.11.012. pages 71

[15] M. Balestrieri, D. Sisti, M. Rocchi, P. Rucci, G. Simon, R. Araya, and G. de Girolamo. Effectiveness of clinical decision support systems and telemedicine on outcomes of depression: a cluster randomized trial in general practice. *Family Practice*, 37(6):731–737, Aug. 2020. doi: 10.1093/fampra/cmaa077. URL https://doi.org/10.1093/fampra/cmaa077. pages 9

[16] T. M. Barros, P. A. Souza Neto, I. Silva, and L. A. Guedes. Predictive models for imbalanced data: A school dropout perspective. *Education Sciences*, 9(4), 2019. ISSN 2227-7102. doi: 10.3390/educsci9040275. URL https://www.mdpi.com/2227-7102/9/4/275. pages 82

[17] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, June 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007735. URL https://doi.org/10.1145/1007730.1007735. pages 84

[18] E. Berner. Diagnostic decision support systems: Why aren' t they used more and what can we do about it? *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2006:1167–8, 02 2006. pages 9

[19] E. S. Berner, editor. *Clinical Decision Support Systems*. Springer New York, 2007. doi: 10.1007/978-0-387-38319-4. URL https://doi.org/10.1007/978-0-387-38319-4. pages 6

[20] E. S. Berner. What can be done to increase the use of diagnostic decision support systems? *Diagnosis*, 1(1):119–123, Jan. 2014.

doi: 10.1515/dx-2013-0014. URL https://doi.org/10.1515/dx-2013-0014. pages 9

[21] A. Bhadoria, K. Sahoo, B. Sahoo, A. Choudhury, N. Sufi, and R. Kumar. Childhood obesity: Causes and consequences. *Journal of Family Medicine and Primary Care*, 4(2):187, 2015. doi: 10.4103/2249-4863.154628. URL https://doi.org/10.4103/2249-4863.154628. pages 75

[22] M. Black, J. Wallace, D. Rankin, P. Carlin, R. Bond, M. Mulvenna, B. Cleland, S. Fischaber, G. Epelde, G. Nikolic, J. Pajula, and R. Connolly. Meaningful integration of data, analytics and services of computer-based medical systems: The MIDAS touch. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, June 2019. doi: 10.1109/cbms.2019.00031. URL https://doi.org/10.1109/cbms.2019.00031. pages 17, 133

[23] J. M. Boden and D. M. Fergusson. Alcohol and depression. *Addiction*, 106(5):906–914, 2011. pages 115

[24] V. Bolón-Canedo and A. Alonso-Betanzos. Ensembles for feature selection: A review and future trends. *Information Fusion*, 52:1–12, 12 2019. ISSN 15662535. doi: 10.1016/j.inffus.2018.11.008. pages 70, 76

[25] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos. Distributed feature selection: An application to microarray data classification. *Applied Soft Computing*, 30:136–150, May 2015. doi: 10.1016/j.asoc.2015.01.035. URL https://doi.org/10.1016/j.asoc.2015.01.035. pages 13, 69, 76, 156

[26] A. L. Booth, E. Abels, and P. McCaffrey. Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Modern Pathology*, 34(3):522–531, Oct. 2020.

doi: 10.1038/s41379-020-00700-x. URL https://doi.org/10.1038/s41379-020-00700-x. pages 4

[27] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug. 1996. doi: 10.1007/bf00058655. URL https://doi.org/10.1007/bf00058655. pages 57

[28] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. pages 58, 60

[29] L. Chang and J. Neu. Early factors leading to later obesity: Interactions of the microbiome, epigenome, and nutrition. *Current Problems in Pediatric and Adolescent Health Care*, 45(5):134–142, May 2015. doi: 10.1016/j.cppeds.2015.03.003. URL https://doi.org/10.1016/j.cppeds.2015.03.003. pages 75

[30] M. Chen-Xu, C. Yokose, S. K. Rai, M. H. Pillinger, and H. K. Choi. Contemporary prevalence of gout and hyperuricemia in the united states and decadal trends: the national health and nutrition examination survey, 2007–2016. *Arthritis & Rheumatology*, 71(6): 991–999, 2019. pages 116

[31] A. K. Cheung, M. Rahman, D. M. Reboussin, T. E. Craven, T. Greene, P. L. Kimmel, W. C. Cushman, A. T. Hawfield, K. C. Johnson, C. E. Lewis, S. Oparil, M. V. Rocco, K. M. Sink, P. K. Whelton, J. T. Wright, J. Basile, S. Beddhu, U. Bhatt, T. I. Chang, G. M. Chertow, M. Chonchol, B. I. Freedman, W. Haley, J. H. Ix, L. A. Katz, A. A. Killeen, V. Papademetriou, A. C. Ricardo, K. Servilla, B. Wall, D. Wolfgram, and J. Yee. Effects of intensive BP control in CKD. *Journal of the American Society of Nephrology*, 28(9):2812–2823, June 2017. doi: 10.1681/asn.2017020148. URL https://doi.org/10.1681/asn.2017020148. pages 92

[32] P. C. Cheung, S. A. Cunningham, K. V. Narayan, and M. R. Kramer. Childhood obesity incidence in the united states: A systematic review.

Childhood Obesity, 12(1):1–11, Feb. 2016. doi: 10.1089/chi.2015.0055. URL https://doi.org/10.1089/chi.2015.0055. pages 49, 50

[33] B. Cleland, J. Wallace, R. R. Bond, M. Black, M. Mulvenna, D. M. Rankin, J. F. Gomes, P. Connolly, G. Epelde, and S. Campbell. Meaningful integration of data analytics and services in MIDAS project: Engaging users in the co-design of a health analytics platform. BCS Learning & Development, July 2018. doi: 10.14236/ewic/hci2018.169. URL https://doi.org/10.14236/ewic/hci2018.169. pages 148

[34] B. Cleland, J. Wallace, R. Bond, S. Muuraiskangas, J. Pajula, G. Epelde, M. Arrúe, R. Álvarez, M. Black, M. D. Mulvenna, D. Rankin, and P. Carlin. Usability evaluation of a co-created big data analytics platform for health policy-making. In *Human Interface and the Management of Information. Visual Information and Knowledge Management*, pages 194–207. Springer International Publishing, 2019. doi: 10.1007/978-3-030-22660-2_13. URL https://doi.org/10.1007/978-3-030-22660-2_13. pages 148

[35] A. J. Collins, S. Li, J. Z. Ma, and C. Herzog. Cardiovascular disease in end-stage renal disease patients. *American Journal of Kidney Diseases*, 38(4):S26–S29, Oct. 2001. doi: 10.1053/ajkd.2001.27392. URL https://doi.org/10.1053/ajkd.2001.27392. pages 83

[36] E. Commission. Rare disease. https://ec.europa.eu/health/non_communicable_diseases/rare_diseases_en. Accessed January 14, 2021. pages 95

[37] J. P. Costa, luka Stopar, F. Fuart, M. Grobelnik, R. Santanam, C. Sun, P. Carlin, M. Black, and J. Wallace. Text mining open datasets to support public health. In *Workshop on Information Technologies and Systems*, December 2017. URL http://oro.open.ac.uk/72252/. pages 145

[38] J. P. Costa, F. Fuart, L. Stopar, M. Grobelnik, D. Mladenić, A. Košmerlj, E. Belayeva, G. Leban, L. Rei, S. Fischaber, and J. Wallace. Health news bias and epidemic intelligence for public health. In *Proceedings of the SiKDD 2019*, volume C, pages 21–24, Oct 2019. pages 146

[39] F. Creed and E. Guthrie. Psychological factors in the irritable bowel syndrome. *Gut*, 28(10):1307–1318, 1987. pages 116

[40] E. F. Davis, M. Lazdam, A. J. Lewandowski, S. A. Worton, B. Kelly, Y. Kenworthy, S. Adwani, A. R. Wilkinson, K. McCormick, I. Sargent, C. Redman, and P. Leeson. Cardiovascular risk factors in children and young adults born to preeclamptic pregnancies: A systematic review. *PEDIATRICS*, 129(6):e1552–e1561, May 2012. doi: 10.1542/peds.2011-3093. URL `https://doi.org/10.1542/peds.2011-3093`. pages 75

[41] K. K. Davison and L. L. Birch. Childhood overweight: a contextual model and recommendations for future research. *Obesity Reviews*, 2 (3):159–171, Aug. 2001. doi: 10.1046/j.1467-789x.2001.00036.x. URL `https://doi.org/10.1046/j.1467-789x.2001.00036.x`. pages 69, 75

[42] J. Dean and S. Ghemawat. MapReduce. *Communications of the ACM*, 51(1):107–113, Jan. 2008. doi: 10.1145/1327452.1327492. URL `https://doi.org/10.1145/1327452.1327492`. pages 56

[43] M. Dehlin, L. Jacobsson, and E. Roddy. Global epidemiology of gout: prevalence, incidence, treatment patterns and risk factors. *Nature Reviews Rheumatology*, 16(7):380–390, 2020. pages 116

[44] J. V. den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst. Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2(10):e267, Sept. 2005. doi: 10.1371/journal.pmed.0020267. URL `https://doi.org/10.1371/journal.pmed.0020267`. pages 35

[45] M. G. Denker and D. L. Cohen. Antihypertensive medications in end-stage renal disease. *Seminars in Dialysis*, 28(4):330–336, Apr. 2015. doi: 10.1111/sdi.12369. URL https://doi.org/10.1111/sdi.12369. pages 83

[46] D. A. Dev, B. A. McBride, B. H. Fiese, B. L. Jones, and o. b. o. t. S. K. R. Hyunkeun Cho. Risk factors for overweight/obesity in preschool children: An ecological approach. *Childhood Obesity*, 9 (5):399–408, Oct. 2013. doi: 10.1089/chi.2012.0150. URL https://doi.org/10.1089/chi.2012.0150. pages 69, 75

[47] J.-F. Diaz-Garelli, E. V. Bernstam, M. Lee, K. O. Hwang, M. H. Rahbar, and T. R. Johnson. DataGauge: A practical process for systematically designing and implementing quality assessments of repurposed clinical data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 7(1), 2019. doi: 10.5334/egems.286. URL https://doi.org/10.5334/egems.286. pages 28

[48] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017. pages 61

[49] O. Dziadkowiec, T. Callahan, M. Ozkaynak, B. Reeder, and J. Welton. Using a data quality framework to clean data extracted from the electronic health record: A case study. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 4(1):11, June 2016. doi: 10.13063/2327-9214.1201. URL https://doi.org/10.13063/2327-9214.1201. pages 10, 11, 28

[50] J. B. Echouffo-Tcheugui and A. P. Kengne. Risk models to predict chronic kidney disease and its progression: A systematic review. *PLoS Medicine*, 9(11):e1001344, Nov. 2012. doi: 10.1371/journal.pmed.1001344. URL https://doi.org/10.1371/journal.pmed.1001344. pages 92

[51] F. F. O. EIZAGUIRRE. Descarga de gráficas y tablas. https://www.fundacionorbegozo.com/el-instituto-de-investigacion-del-crecimiento-y-desarrollo/graficas-y-tablas/. Accessed 07 July 2020. pages 54

[52] A. F. El-Behadli, C. Sharp, S. O. Hughes, E. M. Obasi, and T. A. Nicklas. Maternal depression, stress and feeding styles: towards a framework for theory and research in child obesity. *British Journal of Nutrition*, 113(S1):S55–S71, Jan. 2015. doi: 10.1017/s000711451400333x. URL https://doi.org/10.1017/s000711451400333x. pages 69, 75

[53] P. Embi, A. Jain, J. Clark, and C. Harris. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 231–5, 2005. pages 8

[54] G. Epelde, R. Álvarez, A. Beristain, M. Arrúe, I. Arangoa, and D. Rankin. Enhancing the interactive visualisation of a data preparation tool from in-memory fitting to big data sets. In *Business Information Systems Workshops*, pages 272–284. Springer International Publishing, 2020. doi: 10.1007/978-3-030-61146-0_22. URL https://doi.org/10.1007/978-3-030-61146-0_22. pages 136

[55] S. Eslami, N. F. de Keizer, D. A. Dongelmans, E. de Jonge, M. J. Schultz, and A. Abu-Hanna. Effects of two different levels of computerized decision support on blood glucose regulation in critically ill patients. *International Journal of Medical Informatics*, 81(1):53–60, Jan. 2012. doi: 10.1016/j.ijmedinf.2011.10.004. URL https://doi.org/10.1016/j.ijmedinf.2011.10.004. pages 8

[56] H. Estiri and K. Stephens. DQe-v: A database-agnostic framework for exploring variability in electronic health record data across time and site location. *eGEMs (Generating Evidence & Methods to improve*

patient outcomes), 5(1):3, May 2017. doi: 10.13063/2327-9214.1277. URL https://doi.org/10.13063/2327-9214.1277. pages 28

[57] G. Eysenbach. What is e-health? *Journal of Medical Internet Research*, 3(2):e20, June 2001. doi: 10.2196/jmir.3.2.e20. URL https://doi.org/10.2196/jmir.3.2.e20. pages 2

[58] G. Faraco and C. Iadecola. Hypertension: a harbinger of stroke and dementia. *Hypertension*, 62(5):810–817, 2013. pages 116

[59] S. L. Feder. Data quality in electronic health records research: Quality domains and assessment methods. *Western Journal of Nursing Research*, 40(5):753–766, Jan. 2017. doi: 10.1177/0193945916689084. URL https://doi.org/10.1177/0193945916689084. pages 11, 27, 38

[60] P. Fioretto and A. Solini. Antihypertensive treatment and multifactorial approach for renal protection in diabetes. *Journal of the American Society of Nephrology*, 16(3 suppl 1):S18–S21, Mar. 2005. doi: 10.1681/asn.2004110962. URL https://doi.org/10.1681/asn.2004110962. pages 92

[61] A. J. Flórez, G. Molenberghs, G. Verbeke, M. G. Kenward, P. Mamouris, and B. Vaes. Fast two-stage estimator for clustered count data with overdispersion. *Journal of Statistical Computation and Simulation*, 89(14):2678–2693, June 2019. doi: 10.1080/00949655.2019.1630411. URL https://doi.org/10.1080/00949655.2019.1630411. pages 42

[62] B. F. Fuemmeler, E. Dedert, F. J. McClernon, and J. C. Beckham. Adverse childhood events are associated with obesity and disordered eating: Results from a u.s. population-based survey of young adults. *Journal of Traumatic Stress*, 22(4):329–333, Aug. 2009. doi: 10.1002/jts.20421. URL https://doi.org/10.1002/jts.20421. pages 75

[63] P. A. Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017. pages 104, 117

[64] S. Garavaglia, A. Dun, and B. M. Hill. A smart guide to dummy variables : Four applications and a macro. 1998. pages 74

[65] T. Greenhalgh. How to read a paper: The medline database. *BMJ*, 315(7101):180–183, July 1997. doi: 10.1136/bmj.315.7101.180. URL `https://doi.org/10.1136/bmj.315.7101.180`. pages 145

[66] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3(null):1157–1182, Mar. 2003. ISSN 1532-4435. pages 71

[67] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. *Machine Learning*, 46(1/3):389–422, 2002. doi: 10.1023/a:1012487302797. URL `https://doi.org/10.1023/a:1012487302797`. pages 57

[68] D. Haire-Joshu, R. E. Glasgow, and T. L. Tibbs. Smoking and diabetes. *Diabetes care*, 22(11):1887–1898, 1999. pages 116

[69] R. Hammond, R. Athanasiadou, S. Curado, Y. Aphinyanaphongs, C. Abrams, M. J. Messito, R. Gross, M. Katzow, M. Jay, N. Razavian, and B. Elbel. Predicting childhood obesity using electronic health records and publicly available data. *PLOS ONE*, 14(4):e0215571, Apr. 2019. doi: 10.1371/journal.pone.0215571. URL `https://doi.org/10.1371/journal.pone.0215571`. pages 75

[70] K. HAYRINEN, K. SARANTO, and P. NYKANEN. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*, 77(5):291–304, May 2008. doi: 10.1016/j.ijmedinf.2007.09.001. URL `https://doi.org/10.1016/j.ijmedinf.2007.09.001`. pages 2, 3

[71] F. P. Held, F. Blyth, D. Gnjidic, V. Hirani, V. Naganathan, L. M. Waite, M. J. Seibel, J. Rollo, D. J. Handelsman, R. G. Cumming, et al. Association rules analysis of comorbidity and multimorbidity: The concord health and aging in men project. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 71(5):625–631, 2016. pages 101, 102, 118, 157

[72] P. J. Helmons, B. O. Suijkerbuijk, P. V. N. Panday, and J. G. Kosterink. Drug-drug interaction checking assisted by clinical decision support: a return on investment analysis. *Journal of the American Medical Informatics Association*, 22(4):764–772, Feb. 2015. doi: 10.1093/jamia/ocu010. URL `https://doi.org/10.1093/jamia/ocu010`. pages 8

[73] H. Hemingway, F. W. Asselbergs, J. Danesh, R. Dobson, N. Maniadakis, A. Maggioni, G. J. M. van Thiel, M. Cronin, G. Brobert, P. Vardas, S. D. Anker, D. E. Grobbee, and S. D. and. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *European Heart Journal*, 39(16):1481–1495, Aug. 2017. doi: 10.1093/eurheartj/ehx487. URL `https://doi.org/10.1093/eurheartj/ehx487`. pages 27

[74] Z. M. Hira and D. F. Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015:1–13, June 2015. doi: 10.1155/2015/198363. URL `https://doi.org/10.1155/2015/198363`. pages 51

[75] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, Feb. 1970. doi: 10.1080/00401706.1970.10488634. URL `https://doi.org/10.1080/00401706.1970.10488634`. pages 58

[76] K. Hornik, B. Grün, and M. Hahsler. arules-a computational environment for mining association rules and frequent item sets. *Journal of statistical software*, 14(15):1–25, 2005. pages 105

[77] A. A. Howard, J. H. Arnsten, and M. N. Gourevitch. Effect of alcohol consumption on diabetes mellitus: a systematic review. *Annals of internal medicine*, 140(3):211–219, 2004. pages 115

[78] B. K. Jacobsen and N. A. Aars. Changes in waist circumference and the prevalence of abdominal obesity during 1994–2008 - cross-sectional and longitudinal results from two surveys: the tromsø study. *BMC Obesity*, 3(1), Sept. 2016. doi: 10.1186/s40608-016-0121-5. URL `https://doi.org/10.1186/s40608-016-0121-5`. pages 73

[79] E. Jaurlaritza. Euskadin haurren obesitatea prebenitzeko estrategia. `https://www.euskadi.eus/contenidos/informacion/estrategia_obesidad_infantil/eu_def/adjuntos/Euskadin-Haurren-Obesitatea-Prebenitzeko-Estrategia.pdf`. Accessed 28 June 2021. pages 49

[80] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen, and S. Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5(1):1–10, 2014. pages 101

[81] N. L. Jensen, H. S. Pedersen, M. Vestergaard, S. W. Mercer, C. Glümer, and A. Prior. The impact of socioeconomic status and multimorbidity on mortality: a population-based cohort study. *Clinical epidemiology*, 9:279, 2017. pages 100

[82] P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, May 2012. doi: 10.1038/nrg3208. URL `https://doi.org/10.1038/nrg3208`. pages 2

[83] F. Jiang, S. Zhu, C. Yan, X. Jin, H. Bandla, and X. Shen. Sleep and obesity in preschool children. *The Journal of Pediatrics*, 154(6):

814–818, June 2009. doi: 10.1016/j.jpeds.2008.12.043. URL https://doi.org/10.1016/j.jpeds.2008.12.043. pages 68, 69, 75

[84] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical Care*, 50:S21–S29, July 2012. doi: 10.1097/mlr.0b013e318257dd67. URL https://doi.org/10.1097/mlr.0b013e318257dd67. pages 28

[85] M. Kalliomäki, M. C. Collado, S. Salminen, and E. Isolauri. Early differences in fecal microbiota composition in children may predict overweight. *The American Journal of Clinical Nutrition*, 87(3):534–538, Mar. 2008. doi: 10.1093/ajcn/87.3.534. URL https://doi.org/10.1093/ajcn/87.3.534. pages 75

[86] D. S. Keith, G. A. Nichols, C. M. Gullion, J. B. Brown, and D. H. Smith. Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization. *Archives of Internal Medicine*, 164(6):659, Mar. 2004. doi: 10.1001/archinte.164.6.659. URL https://doi.org/10.1001/archinte.164.6.659. pages 80

[87] B. S. Kennedy. Treating patients with multiple cardiovascular conditions: An analysis of outpatient data in the united states, 2005. *Journal of the National Medical Association*, 100(11):1260–1270, 2008. pages 115

[88] S. Khairat, D. Marc, W. Crosby, and A. A. Sanousi. Reasons for physicians not adopting clinical decision support systems: Critical analysis. *JMIR Medical Informatics*, 6(2):e24, Apr. 2018. doi: 10.2196/medinform.8912. URL https://doi.org/10.2196/medinform.8912. pages 9

[89] Kibana. Kibana. https://www.elastic.co/kibana. pages 145

[90] M. Komaroff, F. Tedla, E. Helzner, and M. A. Joseph. Antihypertensive medications and change in stages of chronic kidney disease. *International Journal of Chronic Diseases*, 2018:1–10, 2018. doi: 10.1155/2018/1382705. URL https://doi.org/10.1155/2018/1382705. pages 83

[91] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6), June 2004. doi: 10.1103/physreve.69.066138. URL https://doi.org/10.1103/physreve.69.066138. pages 57, 59

[92] S. Kumar and A. S. Kelly. Review of childhood obesity. *Mayo Clinic Proceedings*, 92(2):251–265, Feb. 2017. doi: 10.1016/j.mayocp.2016.09.017. URL https://doi.org/10.1016/j.mayocp.2016.09.017. pages 49, 50, 68, 75

[93] R. Kwok, M. Dinh, D. Dinh, and M. Chu. Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: Implementation of a dynamic and integrated electronic decision support system. *Emergency Medicine Australasia*, 21(1):31–37, Feb. 2009. doi: 10.1111/j.1742-6723.2008.01149.x. URL https://doi.org/10.1111/j.1742-6723.2008.01149.x. pages 8

[94] H. Lamberts and M. Wood. The birth of the international classification of primary care (icpc) serendipity at the border of lac leman, 2002. pages 103

[95] M. J. Landray, J. R. Emberson, L. Blackwell, T. Dasgupta, R. Zakeri, M. D. Morgan, C. J. Ferro, S. Vickery, P. Ayrton, D. Nair, R. N. Dalton, E. J. Lamb, C. Baigent, J. N. Townend, and D. C. Wheeler. Prediction of ESRD and death among people with CKD: The chronic renal impairment in birmingham (CRIB) prospective cohort study. *American Journal of Kidney Diseases*, 56(6):1082–1094, Dec. 2010.

doi: 10.1053/j.ajkd.2010.07.016. URL https://doi.org/10.1053/j.ajkd.2010.07.016. pages 14, 81, 92

[96] E. Y. Lau, J. Liu, E. Archer, S. M. McDonald, and J. Liu. Maternal weight gain in pregnancy and risk of obesity among offspring: A systematic review. *Journal of Obesity*, 2014:1–16, 2014. doi: 10.1155/2014/524939. URL https://doi.org/10.1155/2014/524939. pages 75

[97] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. Event registry: Learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 107–110, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327459. doi: 10.1145/2567948.2577024. URL https://doi.org/10.1145/2567948.2577024. pages 145

[98] S.-J. Lee, Z. Xu, T. Li, and Y. Yang. A novel bagging c4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making. *Journal of Biomedical Informatics*, 78:144–155, Feb. 2018. doi: 10.1016/j.jbi.2017.11.005. URL https://doi.org/10.1016/j.jbi.2017.11.005. pages 61

[99] B. Livingstone. Epidemiology of childhood obesity in europe. *European Journal of Pediatrics*, 159(S1):S14–S34, Aug. 2000. doi: 10.1007/pl00014363. URL https://doi.org/10.1007/pl00014363. pages 49

[100] L. Lu, P. Xun, Y. Wan, K. He, and W. Cai. Long-term association between dairy consumption and risk of childhood obesity: a systematic review and meta-analysis of prospective cohort studies. *European Journal of Clinical Nutrition*, 70(4):414–423, Feb. 2016. doi: 10.1038/ejcn.2015.226. URL https://doi.org/10.1038/ejcn.2015.226. pages 49, 50

[101] D. Álvarez Estévez, N. Sánchez-Maroño, A. Alonso-Betanzos, and V. Moret-Bonillo. Reducing dimensionality in a database of sleep eeg

arousals. *Expert Systems with Applications*, 38:7746–7754, 6 2011. ISSN 09574174. doi: 10.1016/j.eswa.2010.12.134. pages 13, 70, 76, 156

[102] T. T. Makovski, S. Schmitz, M. P. Zeegers, S. Stranges, and M. van den Akker. Multimorbidity and quality of life: systematic literature review and meta-analysis. *Ageing research reviews*, 53:100903, 2019. pages 100

[103] G. Maschio, D. Alberti, G. Janin, F. Locatelli, J. F. Mann, M. Motolese, C. Ponticelli, E. Ritz, and P. Zucchelli. Effect of the angiotensin-converting–enzyme inhibitor benazepril on the progression of chronic renal insufficiency. *New England Journal of Medicine*, 334(15):939–945, Apr. 1996. doi: 10.1056/nejm199604113341502. URL `https://doi.org/10.1056/nejm199604113341502`. pages 92

[104] M. Mashoufi, H. Ayatollahi, and D. Khorasani-Zavareh. A review of data quality assessment in emergency medical services. *The Open Medical Informatics Journal*, 12(1):19–32, May 2018. doi: 10.2174/1874431101812010019. URL `https://doi.org/10.2174/1874431101812010019`. pages 27

[105] N. Menachemi and Collum. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, page 47, May 2011. doi: 10.2147/rmhp.s12985. URL `https://doi.org/10.2147/rmhp.s12985`. pages 3

[106] J. Mendlovic, H. Barash, H. Yardeni, Y. Banet-Levi, H. Yonath, and A. Raas-Rothschild. Rare diseases dtc: diagnosis, treatment and care. *Harefuah*, 155(4):241–4, 2016. pages 102

[107] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012. pages 117

[108] Z. Miao, S. Sathyanarayanan, E. Fong, W. Paiva, and D. Delen. An assessment and cleaning framework for electronic health records data. pages 907–912, 2018. pages 10, 11, 28

[109] R. A. Miller. Diagnostic decision support systems. In *Health Informatics*, pages 181–208. Springer International Publishing, 2016. doi: 10.1007/978-3-319-31913-1_11. URL https://doi.org/10.1007/978-3-319-31913-1_11. pages 9

[110] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, Feb. 2019. doi: 10.1016/j.artint.2018.07.007. URL https://doi.org/10.1016/j.artint.2018.07.007. pages 62

[111] S. M. Mohnen, M. J. M. van Oosten, J. Los, M. J. H. Leegte, K. J. Jager, M. H. Hemmelder, S. J. J. Logtenberg, V. S. Stel, L. H. van Roijen, and G. A. de Wit. Healthcare costs of patients on different renal replacement modalities – analysis of dutch health insurance claims data. *PLOS ONE*, 14(8):e0220800, Aug. 2019. doi: 10.1371/journal.pone.0220800. URL https://doi.org/10.1371/journal.pone.0220800. pages 80

[112] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020. pages 5

[113] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos. Centralized vs. distributed feature selection methods based on data complexity measures. *Knowledge-Based Systems*, 117:27–45, Feb. 2017. doi: 10.1016/j.knosys.2016.09.022. URL https://doi.org/10.1016/j.knosys.2016.09.022. pages 13, 69, 76, 156

[114] E. Murphy. Clinical decision support: Effectiveness in improving quality processes and clinical outcomes and factors that may influence success. *The Yale journal of biology and medicine*, 87:187–197, 06 2014. pages 9

[115] M. Nagai, S. Hoshide, and K. Kario. Hypertension and dementia. *American journal of hypertension*, 23(2):116–124, 2010. pages 116

[116] S. K. Ng, R. Tawiah, M. Sawyer, and P. Scuffham. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. *International journal of epidemiology*, 47(5):1687–1704, 2018. pages 101

[117] M. Nielen, R. Davids, M. Gommer, R. Poos, and R. Verheij. Berekening morbiditeitscijfers op basis van nivel zorgregistraties eerste lijn. `https://www.volksgezondheidenzorg.info/sites/default/files/documentatie_episodeconstructie_nivel_1juli2016_definitief.pdf`, jul 2016. Accessed January 10, 2020. pages 104

[118] C. Njuguna, M. Vandi, M. Mugagga, J. Kanu, E. Liyosi, A. Chimbaru, W. Gachari, V. Caulker, S. Sesay, J. Githuku, Z. Yoti, A. A. Yahaya, A. Talisuna, and I. S. Fall. Institutionalized data quality assessments: a critical pathway to improving the accuracy of integrated disease surveillance data in sierra leone. *BMC Health Services Research*, 20(1), Aug. 2020. doi: 10.1186/s12913-020-05591-x. URL `https://doi.org/10.1186/s12913-020-05591-x`. pages 11, 28

[119] M. C. Ochoa, M. J. Moreno-Aliaga, M. A. Martínez-González, J. A. Martínez, and A. Marti. Predictor factors for childhood obesity in a spanish case-control study. *Nutrition*, 23(5):379–384, May 2007. doi: 10.1016/j.nut.2007.02.004. URL `https://doi.org/10.1016/j.nut.2007.02.004`. pages 71

[120] D. of General Practice. Intego-project. `http://www.intego.be`. Accessed March 30, 2020. pages 19, 102

[121] N. D. of Health and H. Services. Implementing clinical decision support systems. `shorturl.at/glEW1`. Accessed April 10, 2021. pages 9

[122] B. Olaya, M. V. Moneta, F. F. Caballero, S. Tyrovolas, I. Bayes, J. L. Ayuso-Mateos, and J. M. Haro. Latent class analysis of multimorbidity patterns and associated outcomes in spanish older adults: a prospective cohort study. *BMC geriatrics*, 17(1):1–10, 2017. pages 101

[123] W. H. Organization. Data and statistics. `http://www.euro.who.int/en/health-topics/noncommunicable-diseases/obesity/data-and-statistics`. Accessed 07 July 2020. pages 49

[124] Osakidetza. `https://www.osakidetza.euskadi.eus/`. Accessed 07 July 2020. pages 53

[125] X. Pang, C. B. Forrest, F. Lê-Scherban, and A. J. Masino. Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, 150:104454, June 2021. doi: 10.1016/j.ijmedinf.2021.104454. URL `https://doi.org/10.1016/j.ijmedinf.2021.104454`. pages 49

[126] R. B. Parikh, M. Kakad, and D. W. Bates. Integrating predictive analytics into high-value care. *JAMA*, 315(7):651, Feb. 2016. doi: 10.1001/jama.2015.19417. URL `https://doi.org/10.1001/jama.2015.19417`. pages 4

[127] S. Park, D. Choi, M. Kim, W. Cha, C. Kim, and I.-C. Moon. Identifying prescription patterns with a topic model of diseases and medications. *Journal of biomedical informatics*, 75:35–47, 2017. pages 101, 114

[128] J. Pei, J. Han, R. Mao, et al. Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, volume 4, pages 21–30, 2000. pages 119

[129] N. Peiffer-Smadja, T. Rawson, R. Ahmad, A. Buchard, P. Georgiou, F.-X. Lescure, G. Birgand, and A. Holmes. Machine learning for

clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5): 584–595, May 2020. doi: 10.1016/j.cmi.2019.09.009. URL https://doi.org/10.1016/j.cmi.2019.09.009. pages 8

[130] M. J. Pencina and E. D. Peterson. Moving from clinical trials to precision medicine. *JAMA*, 315(16):1713, Apr. 2016. doi: 10.1001/jama.2016.4839. URL https://doi.org/10.1001/jama.2016.4839. pages 4

[131] N. Pérez-Farinós, A. M. López-Sobaler, M. Á. D. Re, C. Villar, E. Labrado, T. Robledo, and R. M. Ortega. The ALADINO study: A national study of prevalence of overweight and obesity in spanish children in 2011. *BioMed Research International*, 2013:1–7, 2013. doi: 10.1155/2013/163687. URL https://doi.org/10.1155/2013/163687. pages 70

[132] D. Peteiro-Barral and B. Guijarro-Berdiñas. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, 2 (1):1–11, Nov. 2012. doi: 10.1007/s13748-012-0035-5. URL https://doi.org/10.1007/s13748-012-0035-5. pages 76

[133] H. T. T. Phan, F. Borca, D. Cable, J. Batchelor, J. H. Davies, and S. Ennis. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. *Scientific Reports*, 10(1), June 2020. doi: 10.1038/s41598-020-66925-7. URL https://doi.org/10.1038/s41598-020-66925-7. pages 12, 28

[134] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, pages 229–238, 1991. pages 101

[135] J. Pita Costa, M. Grobelnik, F. Fuart, L. Stopar, G. Epelde, S. Fischaber, P. Poliwoda, D. Rankin, J. Wallace, M. Black, R. Bond, M.

Mulvenna, D. Weston, P. Carlin, R. Bilbao, G. Nikolic, X. Shi, and B. De Moor. Meaningful big data integration for a global covid-19 strategy. *IEEE Computational Intelligence Magazine*, 15(4):51–61, 2020. doi: 10.1109/MCI.2020.3019898. pages 150

[136] E. Poskitt. Defining childhood obesity the relative body mass index (bmi).european childhood obesity group. *Acta Paediatrica*, 84(8), 1995. pages 54

[137] G. V. Pottelbergh, S. Bartholomeeusen, F. Buntinx, and J. Degryse. The evolution of renal function and the incidence of end-stage renal disease in patients aged over 50 years. *Nephrology Dialysis Transplantation*, 27(6):2297–2303, Nov. 2011. doi: 10.1093/ndt/gfr659. URL `https://doi.org/10.1093/ndt/gfr659`. pages 81, 83

[138] G. V. Pottelbergh, P. Mamouris, N. Opdeweegh, B. Vaes, G. Goderis, and M. V. D. Akker. Is there a correlation between an eGFR slope measured over a 5-year period and incident cardiovascular events in the following 5 years among a flemish general practice population: a retrospective cohort study. *BMJ Open*, 8(11):e023594, Nov. 2018. doi: 10.1136/bmjopen-2018-023594. URL `https://doi.org/10.1136/bmjopen-2018-023594`. pages 42

[139] C. Pozzi, P. Bolasco, G. Fogazzi, S. Andrulli, P. Altieri, C. Ponticelli, and F. Locatelli. Corticosteroids in IgA nephropathy: a randomised controlled trial. *The Lancet*, 353(9156):883–887, Mar. 1999. doi: 10.1016/s0140-6736(98)03563-6. URL `https://doi.org/10.1016/s0140-6736(98)03563-6`. pages 92

[140] J. B. Pryor, B. R. Weber, J. V. Weber, J. B. Lockridge, and A. J. Olyaei. Lipid-lowering agents for the treatment of hyperlipidemia in patients with chronic kidney disease and end-stage renal disease on dialysis: a review. *Drugs & Therapy Perspectives*, 35(9):431–

441, June 2019. doi: 10.1007/s40267-019-00646-4. URL https://doi.org/10.1007/s40267-019-00646-4. pages 84

[141] G. Ramkumar, S. Ranka, and S. Tsur. Weighted association rules: Model and algorithm. In *Proc. ACM SIGKDD*, pages 661–666. Citeseer, 1998. pages 101, 105

[142] S. Ramírez-Gallego, H. Mouriño-Talín, D. Martínez-Rego, V. Bolón-Canedo, J. M. Benítez, A. Alonso-Betanzos, and F. Herrera. An information theory-based feature selection framework for big data under apache spark. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9):1441–1453, 2018. doi: 10.1109/TSMC.2017.2670926. pages 13, 69, 156

[143] C. L. Ramspek, Y. de Jong, F. W. Dekker, and M. van Diepen. Towards the best kidney failure prediction tool: a systematic review and selection aid. *Nephrology Dialysis Transplantation*, 35(9):1527–1538, Mar. 2019. doi: 10.1093/ndt/gfz018. URL https://doi.org/10.1093/ndt/gfz018. pages 14, 81, 92

[144] C. E. R. Results. Midas framework user guide. https://cordis.europa.eu/project/id/727721/results. pages 149

[145] P. C. Robinson, W. J. Taylor, and N. Dalbeth. An observational study of gout prevalence and quality of care in a national australian general practice population. *The Journal of rheumatology*, 42(9):1702–1707, 2015. pages 116

[146] M. Robnik-Šikonja and M. Bohanec. Perturbation-based explanations of prediction models. In *Human and Machine Learning*, pages 159–175. Springer International Publishing, 2018. doi: 10.1007/978-3-319-90403-0_9. URL https://doi.org/10.1007/978-3-319-90403-0_9. pages 62

[147] E. G. Ross, N. H. Shah, R. L. Dalman, K. T. Nead, J. P. Cooke, and N. J. Leeper. The use of machine learning for the identification of peripheral artery disease and future mortality risk. *Journal of Vascular Surgery*, 64(5):1515–1522.e3, Nov. 2016. doi: 10.1016/j.jvs.2016.04.026. URL `https://doi.org/10.1016/j.jvs.2016.04.026`. pages 4

[148] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. doi: 10.1038/s42256-019-0048-x. URL `https://doi.org/10.1038/s42256-019-0048-x`. pages 4

[149] P. Ryan, Y. Zhang, F. Liu, J. Gao, J. Bigger, G. Hripcsak, C. Weng, and Y. Li. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Applied Clinical Informatics*, 05(02):463–479, 2014. doi: 10.4338/aci-2013-12-ra-0105. URL `https://doi.org/10.4338/aci-2013-12-ra-0105`. pages 3

[150] R. Á. Sánchez, A. B. Iraola, G. E. Unanue, and P. Carlin. TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. *Computer Methods and Programs in Biomedicine*, 181:104824, Nov. 2019. doi: 10.1016/j.cmpb.2018.12.029. URL `https://doi.org/10.1016/j.cmpb.2018.12.029`. pages 136

[151] A. Sarker and G. Gonzalez-Hernandez. An unsupervised and customizable misspelling generator for mining noisy health-related text sources. *Journal of Biomedical Informatics*, 88:98–107, Dec. 2018. doi: 10.1016/j.jbi.2018.11.007. URL `https://doi.org/10.1016/j.jbi.2018.11.007`. pages 33

[152] I. Schäfer, H. Kaduszkiewicz, H.-O. Wagner, G. Schön, M. Scherer, and H. van den Bussche. Reducing complexity: a visualisation of multimorbidity by combining disease clusters and triads. *BMC Public Health*, 14(1):1–14, 2014. pages 101

[153] C. M. Schooling and H. E. Jones. Clarifying questions about "risk factors": predictors versus explanation. *Emerging Themes in Epidemiology*, 15(1), Aug. 2018. doi: 10.1186/s12982-018-0080-z. URL https://doi.org/10.1186/s12982-018-0080-z. pages 71

[154] K. U. Schulz and S. Mihov. Fast string correction with levenshtein automata. *International Journal on Document Analysis and Recognition*, 5(1):67–85, Nov. 2002. doi: 10.1007/s10032-002-0082-8. URL https://doi.org/10.1007/s10032-002-0082-8. pages 33

[155] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos. Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowledge-Based Systems*, 118:124–139, Feb. 2017. doi: 10.1016/j.knosys.2016.11.017. URL https://doi.org/10.1016/j.knosys.2016.11.017. pages 13, 69, 156

[156] M. Sekine, T. Yamagami, K. Handa, T. Saito, S. Nanri, K. Kawaminami, N. Tokui, K. Yoshida, and S. Kagamimori. A dose-response relationship between short sleeping hours and childhood obesity: results of the toyama birth cohort study. *Child: Care, Health and Development*, 28(2):163–170, Mar. 2002. doi: 10.1046/j.1365-2214.2002.00260.x. URL https://doi.org/10.1046/j.1365-2214.2002.00260.x. pages 68, 69, 75

[157] H. D. Sesso, N. R. Cook, J. E. Buring, J. E. Manson, and J. M. Gaziano. Alcohol consumption and the risk of hypertension in women and men. *Hypertension*, 51(4):1080–1087, 2008. pages 115

[158] A. M. Shahsavarani, E. Azad Marz Abadi, M. Hakimi Kalkhoran, S. Jafari, and S. Qaranli. Clinical decision support systems (cdsss): State of the art review of literature. *International Journal of Medical Reviews*, 2(4):299–308, 2015. ISSN 2345-525X. URL http://www.ijmedrev.com/article_68717.html. pages 6

[159] K. Shailaja, B. Seetharamulu, and M. A. Jabbar. Machine learning in healthcare: A review. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 910–914, 2018. doi: 10.1109/ICECA.2018.8474918. pages 4

[160] X. Shi. Mining public health data records using various machine learning methods. Master's thesis, KU Leuven, 2017. pages 20, 21

[161] F. A. Sonnenberg and J. R. Beck. Markov models in medical decision making. 13(4):322–338, Dec. 1993. doi: 10.1177/0272989x9301300409. URL `https://doi.org/10.1177/0272989x9301300409`. pages 117

[162] M. D. Soufi, T. Samad-Soltani, S. S. Vahdati, and P. Rezaei-Hachesu. Decision support system for triage management: A hybrid approach using rule-based reasoning and fuzzy logic. *International Journal of Medical Informatics*, 114:35–44, June 2018. doi: 10.1016/j.ijmedinf.2018.03.008. URL `https://doi.org/10.1016/j.ijmedinf.2018.03.008`. pages 6

[163] L. Souza-Pereira, N. Pombo, S. Ouhbi, V. Felizardo, and N. Garcia. Clinical decision support systems for chronic diseases: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 195:105565, Oct. 2020. doi: 10.1016/j.cmpb.2020.105565. URL `https://doi.org/10.1016/j.cmpb.2020.105565`. pages 8

[164] S. Speedie, G. Simon, V. Kumar, B. Westra, and S. Johnson. Application of an ontology for characterizing data quality for a secondary use of EHR data. *Applied Clinical Informatics*, 07(01):69–88, Jan. 2016. doi: 10.4338/aci-2015-08-ra-0107. URL `https://doi.org/10.4338/aci-2015-08-ra-0107`. pages 11, 28

[165] P. Srinivasan and B. Libbus. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20(Suppl 1):i290–i296, July 2004. doi: 10.1093/bioinformatics/bth914. URL `https://doi.org/10.1093/bioinformatics/bth914`. pages 145

[166] V. S. Stel, K. Brück, S. Fraser, C. Zoccali, Z. A. Massy, and K. J. Jager. International differences in chronic kidney disease prevalence: a key public health and epidemiologic research issue. *Nephrology Dialysis Transplantation*, 32(suppl_2):ii129–ii135, Feb. 2017. doi: 10.1093/ndt/gfw420. URL https://doi.org/10.1093/ndt/gfw420. pages 80

[167] L. A. Stevens, J. Coresh, T. Greene, and A. S. Levey. Assessing kidney function — measured and estimated glomerular filtration rate. *New England Journal of Medicine*, 354(23):2473–2483, June 2006. doi: 10.1056/nejmra054415. URL https://doi.org/10.1056/nejmra054415. pages 83

[168] P. E. Stevens. Evaluation and management of chronic kidney disease: Synopsis of the kidney disease: Improving global outcomes 2012 clinical practice guideline. *Annals of Internal Medicine*, 158(11):825, June 2013. doi: 10.7326/0003-4819-158-11-201306040-00007. URL https://doi.org/10.7326/0003-4819-158-11-201306040-00007. pages 81

[169] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10 (5), June 2020. doi: 10.1002/widm.1379. URL https://doi.org/10.1002/widm.1379. pages 5

[170] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), Jan. 2007. doi: 10.1186/1471-2105-8-25. URL https://doi.org/10.1186/1471-2105-8-25. pages 61

[171] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*,

302(5643):249–255, 2003. ISSN 0036-8075. doi: 10.1126/science.1087447. URL https://science.sciencemag.org/content/302/5643/249. pages 13, 70, 77, 156

[172] M. Sud, N. Tangri, A. Levin, M. Pintilie, A. S. Levey, and D. M. Naimark. CKD stage at nephrology referral and factors influencing the risks of ESRD and death. *American Journal of Kidney Diseases*, 63(6):928–936, June 2014. doi: 10.1053/j.ajkd.2013.12.008. URL https://doi.org/10.1053/j.ajkd.2013.12.008. pages 83

[173] D. Sun and D. Zhang. Bagging constraint score for feature selection with pairwise constraints. *Pattern Recognition*, 43(6):2106–2118, June 2010. doi: 10.1016/j.patcog.2009.12.011. URL https://doi.org/10.1016/j.patcog.2009.12.011. pages 61

[174] K. Sun and F. Bai. Mining weighted association rules without preassigned weights. 20(4):489–495, Apr. 2008. doi: 10.1109/tkde.2007.190723. URL https://doi.org/10.1109/tkde.2007.190723. pages 119

[175] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1), Feb. 2020. doi: 10.1038/s41746-020-0221-y. URL https://doi.org/10.1038/s41746-020-0221-y. pages 6, 8, 9

[176] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. W. Sjoding, and J. Wiens. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, Oct. 2020. doi: 10.1093/jamia/ocaa139. URL https://doi.org/10.1093/jamia/ocaa139. pages 28

[177] E. M. Taveras. Crossing growth percentiles in infancy and risk of obesity in childhood. *Archives of Pediatrics & Adolescent Medicine*,

165(11):993, Nov. 2011. doi: 10.1001/archpediatrics.2011.167. URL https://doi.org/10.1001/archpediatrics.2011.167. pages 75

[178] A. L. Terry, M. Stewart, S. Cejic, J. N. Marshall, S. de Lusignan, B. M. Chesworth, V. Chevendra, H. Maddocks, J. Shadd, F. Burge, and A. Thind. A basic model for assessing primary health care electronic medical record data quality. *BMC Medical Informatics and Decision Making*, 19(1), Feb. 2019. doi: 10.1186/s12911-019-0740-0. URL https://doi.org/10.1186/s12911-019-0740-0. pages 11, 27, 38

[179] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178. pages 58

[180] K. A. Timmins, M. A. Green, D. Radley, M. A. Morris, and J. Pearce. How has big data contributed to obesity research? a review of the literature. *International Journal of Obesity*, 42(12):1951–1962, July 2018. doi: 10.1038/s41366-018-0153-7. URL https://doi.org/10.1038/s41366-018-0153-7. pages 49

[181] M. Tonelli, N. Wiebe, B. Culleton, A. House, C. Rabbat, M. Fok, F. McAlister, and A. X. Garg. Chronic kidney disease and mortality risk: A systematic review. *Journal of the American Society of Nephrology*, 17(7):2034–2047, 2006. doi: 10.1681/asn.2005101085. URL https://doi.org/10.1681/asn.2005101085. pages 80

[182] C. T. Tran, M. Zhang, P. Andreae, and B. Xue. Bagging and feature selection for classification with incomplete data. In *Applications of Evolutionary Computation*, pages 471–486. Springer International Publishing, 2017. doi: 10.1007/978-3-319-55849-3_31. URL https://doi.org/10.1007/978-3-319-55849-3_31. pages 61

[183] R. Tripathi and S. Mehrotra. Irritable bowel syndrome and its psychological management. *Industrial psychiatry journal*, 24(1):91, 2015. pages 116

[184] C. Truyers, G. Goderis, H. Dewitte, M. vanden Akker, and F. Buntinx. The intego database: background, methods and basic results of a flemish general practice-based continuous morbidity registration project. *BMC Medical Informatics and Decision Making*, 14(1), June 2014. doi: 10.1186/1472-6947-14-48. URL `https://doi.org/10.1186/1472-6947-14-48`. pages 20, 29, 82, 102, 103

[185] I. T. Union. Implementing e-health in developing countries: Guidance and principles. `https://www.itu.int/ITU-D/cyb/app/docs/e-Health_prefinal_15092008.PDF`, Sep 2008. Accessed March 18, 2021. pages 2

[186] B. Vaes, E. Beke, C. Truyers, S. Elli, F. Buntinx, J. Y. Verbakel, G. Goderis, and G. V. Pottelbergh. The correlation between blood pressure and kidney function decline in older people: a registry-based cohort study. *BMJ Open*, 5(6):e007571, June 2015. doi: 10.1136/bmjopen-2015-007571. URL `https://doi.org/10.1136/bmjopen-2015-007571`. pages 83

[187] F. J. Valverde-Albacete and C. Peláez-Moreno. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE*, 9(1):e84217, Jan. 2014. doi: 10.1371/journal.pone.0084217. URL `https://doi.org/10.1371/journal.pone.0084217`. pages 81

[188] M. van den Akker, F. Buntinx, and J. A. Knottnerus. Comorbidity or multimorbidity: what's in a name? a review of literature. *The European journal of general practice*, 2(2):65–70, 1996. pages 100

[189] M. van den Akker, B. Vaes, G. Goderis, G. Van Pottelbergh, T. De Burghgraeve, and S. Henrard. Trends in multimorbidity and

polypharmacy in the flemish-belgian population between 2000 and 2015. *PloS one*, 14(2):e0212046, 2019. pages 100

[190] M. H. C. van Rijn, M. van de Luijtgaarden, A. D. van Zuilen, P. J. Blankestijn, J. F. M. Wetzels, T. P. A. Debray, and J. A. J. G. van den Brand. Prognostic models for chronic kidney disease: a systematic review and external validation. *Nephrology Dialysis Transplantation*, Oct. 2020. doi: 10.1093/ndt/gfaa155. URL `https://doi.org/10.1093/ndt/gfaa155`. pages 14, 81

[191] J. Varghese, M. Kleine, S. I. Gessner, S. Sandmann, and M. Dugas. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. *Journal of the American Medical Informatics Association*, 25(5):593–602, Sept. 2017. doi: 10.1093/jamia/ocx100. URL `https://doi.org/10.1093/jamia/ocx100`. pages 9

[192] K. Wang and M.-Y. T. Su. Item selection by "hub-authority" profit ranking. ACM Press, 2002. doi: 10.1145/775047.775144. URL `https://doi.org/10.1145/775047.775144`. pages 119

[193] M. Warner, A. Wesselink, K. G. Harley, A. Bradman, K. Kogut, and B. Eskenazi. Prenatal exposure to dichlorodiphenyltrichloroethane and obesity at 9 years of age in the CHAMACOS study cohort. *American Journal of Epidemiology*, 179(11):1312–1322, Apr. 2014. doi: 10.1093/aje/kwu046. URL `https://doi.org/10.1093/aje/kwu046`. pages 75

[194] N. G. Weiskopf and C. Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, Jan. 2013. doi: 10.1136/amiajnl-2011-000681. URL `https://doi.org/10.1136/amiajnl-2011-000681`. pages 11, 27, 37

[195] N. G. Weiskopf, S. Bakken, G. Hripcsak, and C. Weng. A data quality assessment guideline for electronic health record data reuse. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 5(1): 14, Sept. 2017. doi: 10.5334/egems.218. URL `https://doi.org/10.5334/egems.218`. pages 11, 28

[196] Wikipedia. Accuracy paradox. shorturl.at/mABEP. Accessed December 21st, 2020. pages 81

[197] M. G. Wong, V. Perkovic, J. Chalmers, M. Woodward, Q. Li, M. E. Cooper, P. Hamet, S. Harrap, S. Heller, S. MacMahon, G. Mancia, M. Marre, D. Matthews, B. Neal, N. Poulter, A. Rodgers, B. Williams, and S. Zoungas. Long-term benefits of intensive glucose control for preventing end-stage kidney disease: ADVANCE-ON. *Diabetes Care*, 39(5):694–700, Mar. 2016. doi: 10.2337/dc15-2322. URL `https://doi.org/10.2337/dc15-2322`. pages 92

[198] X. Xu, G. D. Mishra, and M. Jones. Evidence on multimorbidity from definition to intervention: an overview of systematic reviews. *Ageing research reviews*, 37:53–68, 2017. pages 100

[199] Z. B. Yu, S. P. Han, G. Z. Zhu, C. Zhu, X. J. Wang, X. G. Cao, and X. R. Guo. Birth weight and subsequent risk of obesity: a systematic review and meta-analysis. *Obesity Reviews*, 12(7):525–542, Mar. 2011. doi: 10.1111/j.1467-789x.2011.00867.x. URL `https://doi.org/10.1111/j.1467-789x.2011.00867.x`. pages 75

[200] M. Zaki. Scalable algorithms for association mining. 12(3):372–390, 2000. doi: 10.1109/69.846291. URL `https://doi.org/10.1109/69.846291`. pages 119

[201] D. T. Zemedikun, L. J. Gray, K. Khunti, M. J. Davies, and N. N. Dhalwani. Patterns of multimorbidity in middle-aged and older adults: an analysis of the uk biobank data. In *Mayo Clinic Proceedings*, volume 93, pages 857–866. Elsevier, 2018. pages 101, 102, 104, 114, 118, 157

[202] E. Zitt, C. Pscheidt, H. Concin, R. Kramar, K. Lhotta, and G. Nagel. Anthropometric and metabolic risk factors for ESRD are disease-specific: Results from a large population-based cohort study in austria. *PLOS ONE*, 11(8):e0161376, Aug. 2016. doi: 10.1371/journal.pone.0161376. URL `https://doi.org/10.1371/journal.pone.0161376`. pages 83

List of publications

International journal papers (published)

- Shi X, Nikolic G, Van Pottelbergh G, van den Akker M, Vos R, De Moor B. (2021). Development of multimorbidity over time: an analysis of Belgium primary care data using Markov chains and Weighted Association Rules Mining. *Journals Of Gerontology:Series A* 76(7):1234-1241. doi: 10.1093/gerona/glaa278

- Shi X, Nikolic G, Epelde G, Arrue M, Bidaurrazaga Van-Dierdonck J,Bilbao R, De Moor B. (2021) An Ensemble-Based Feature Selection Framework to Select Risk Factors of Childhood Obesity for Policy Decision Making. *BMC Medical Informatics and Decision Making* 21:222. https://doi.org/10.1186/s12911-021-01580-0

- Shi X, Prins C, Van Pottelbergh G, Mamouris P, Vaes B, De Moor B. (2021) An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge. *BMC Medical Informatics and Decision Making* 21:267. https://doi.org/10.1186/s12911-021-01630-7

- Costa JP, Grobelnik M, Fuart F, Stopar L, Epelde G, Fischaber S, Poliwoda P, Rankin D, Wallace J, Black M, Bond R, Mulvenna M, Weston D, Carlin P, Bilbao R, Nikolic G, Shi X, De Moor B,

Pikkarainen M, Paakkonen J, Staines A, Connolly R, Davis P. (2020) Meaningful Big Data Integration for a Global COVID-19 Strategy. *IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE* 15(4):51-61. doi: 10.1109/MCI.2020.3019898

- Epelde G, Beristain A, Alvarez R, Arrue M, Ezkerra I, Belar O, Bilbao R, Nikolic G, Shi X, De Moor B, Mulvenna M. (2020) Quality of data measurements in the Big Data era - Lessons learned from MIDAS project. *IEEE Instrumentation and Measurement Magazine* 23(7):18-24. doi: 10.1109/MIM.2020.9234761

International journal papers (submitted)

- Shi X, Qu T, Van Pottelbergh G, van den Akker M, De Moor B. (2021) A prognostic model for End-Stage Renal Disease: an analysis of Belgium primary care data.
  Submitted to Frontiers in Medicine.

- Shi X, Nikolic G, Fischaber S, Black M, Rankin D, Epelde G, Beristain A, Alvarez R, Arrue M, Pita Costa J, Grobelnik M, Stopar L, Pajula J, Umer A, Poliwoda P,Wallace J, De Moor B. (2021) System Architecture of A European Platform for Health Policy Decision Making: MIDAS.
  Submitted to Journal of Medical Internet Research.

Conference oral presentation

- Shi X, De Moor B. (2019) Beyond descriptive analysis – a use case of children care system in Northern Ireland. In MIDAS workshop. San Sebastian, Spain.

- Nikolic G, Shi X, De Moor B. (2020) Meaningful integration of data, analytics and services – Pilot study. In Big Data for Health Policy Symposium. Londonderry, the U.K.



FACULTY OF ENGINEERING SCIENCE

DEPARTMENT OF ELECTRICAL ENGINEERING

STADIUS

Kasteelpark Arenberg 10, bus 2446

B-3001 Leuven

xi.shi@esat.kuleuven.be

http://esat.kuleuven.be/stadius