TÉCNICO
LISBOA

Identifying Subgroups in Heart Failure Patients with Multimorbidity by Cluster Analysis

Catarina Patrício Mendes Martins

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Andreia Sofia Monteiro Teixeira
Prof. Mário Jorge Costa Gaspar da Silva

Examination Committee

Chairperson: Prof. Rita Homem de Gouveia Costanzo Nunes
Supervisor: Prof. Mário Jorge Costa Gaspar da Silva
Member of the Committee: Prof. Dulce Alves Brito
Prof. Maria do Rosário de Oliveira Silva

October 2021

Preface

The work presented in this thesis was performed at Hospital da Luz Lisboa (Lisbon, Portugal), during the period March-October 2021, under the supervision of Prof. Andreia Sofia Teixeira, within the scope of the IntelligentCare project LISBOA-01-0247-FEDER-045948 that is co-financed by the 262ERDF/LISBOA2020 and by FCT under CMU-Portugal and by FCT under project UIDB/50021/2020. The thesis was co-supervised at Instituto Superior Técnico by Prof. Mário Gaspar da Silva.

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would first like to thank my supervisors Prof. Mário Silva and Prof. Sofia Teixeira for all the support and motivation during these past months. Even with a late start, they always pushed me to go one step further and were always available for guidance and advice.

I would also like to express my gratitude to Dr. Pedro Sarmento, for guiding me through the complex world of Heart Failure, and for the important discussions and feedback, and to Dr. Bernardo Neves, for the availability and relevant insights. A thank you is also due to Miguel Froes, for being the best data manager and ensuring everything was going smoothly.

This dissertation also marks the end of a chapter. And since ”um curso não se faz sozinho”, I would like to thank those that made this chapter a very happy one. To Margas and Maria, my forever lab partners, to Pipa, for matching my stress levels, to Carol, for all the train and car rides, to Lu, for being my thesis partner and for all the voice messages, to Teresa, for always having the door open, to Maria, for always checking in, and to João and Gonçalo, for always giving the feeling that everything will work out. Thank you for your friendship, for the company in the (way too) long hours in the libraries, and for the churrascos, trips, dinners and all the moments in between.

Another big thank you goes to Clarisse and Maria, for always finding the time for a cup of tea between our not-so-easy-to-match schedules, and for being there every step of the way.

To Rafa, that had the (un)pleasure of following this thesis almost paragraph by paragraph and was still always available to hear my latest (always justified) drama. Thank you for believing in me and supporting me in more ways than I could put into words.

Finally, to my family, Mum, Dad and Miguel. Thank you for encouraging me to always do my best, believing in me and, very rarely (arguable), putting up with my bad humour. Thank you for providing me with all the tools I could need to grow and for being the best support system I could ask for.

Resumo

Esta dissertação propõe um workflow para a identificação e caracterização de subgrupos clinicamente significativos de doentes com Insuficiência Cardíaca (IC), utilizando dados dos Registos Clínicos Eletrónicos. A multimorbilidade, definida como a co-ocorrência de duas ou mais doenças crónicas, representa um grande encargo para os sistemas de saúde. Contudo, existe ainda uma falta de conhecimento sobre os doentes com multimorbilidade, quais as interacções mais comuns entre doenças, os principais fatores de risco, e a resposta aos regimes terapêuticos nestes doentes. Isto é particularmente relevante em condições complexas e heterogéneas, tais como Insuficiência Cardíaca. Foi realizada uma análise de clustering a dados clínicos de pacientes com IC do Hospital da Luz Lisboa e os clusters foram caracterizados em termos de dados demográficos, comorbidades, análises laboratoriais, prescrições e consultas. Foi também realizada uma análise de sobrevivência para testar a associação dos clusters com outcomes, tais como, hospitalização e admissão de urgência. O workflow identificou quatro clusters com diferenças significativas nas variáveis clínicas analisadas. Os resultados da análise de sobrevivência mostraram que os clusters tinham diferentes níveis de risco para hospitalizações e admissões de urgências e que os níveis de risco estavam de acordo com os perfis traçados para clusters. Os resultados evidenciam um elevado grau de heterogeneidade dentro dos doentes com IC e como uma melhor caracterização dos subgrupos de doentes pode ser relevante para a estratificação do risco clínico.

Palavras-chave: Multimorbilidade, Registos Clínicos Eletrónicos, Insuficiência Cardíaca, Fenotipagem, Clustering

Abstract

This dissertation proposes a workflow for the identification and characterisation of clinically significant Heart Failure (HF) patients subgroups with multimorbidity using data from Electronic Health Records. Multimorbidity, defined as the co-occurrence of two or more chronic conditions, represents a heavy burden for healthcare systems. However, there is still a lack of knowledge about patients with multimorbidity, the most common disease interactions, risk factors and treatment response. This is particularly relevant in complex and heterogeneous conditions such as HF. Using clinical data from Hospital da Luz Lisboa, we conducted a clustering analysis of HF patients and characterised the clusters in terms of demographics, comorbidities, laboratory values, medical prescriptions and medical appointments. This was followed by a survival analysis to identify associations between clusters and outcomes such as hospitalisation and emergency admission. The workflow identified four distinct clusters with significant differences in the clinical variables analysed. The survival analysis showed differential associations for hospitalisation and emergency admission risks between clusters. Moreover, the risk associations were in agreement with the cluster profiles. These results evidence a high degree of disease heterogeneity within HF patients and how an improved characterisation of patient subgroups can be relevant for clinical risk stratification.

Keywords: Multimorbidity, Heart Failure, Electronic Health Records, Phenotyping, Clustering

Contents

Acknowledgments . . . vii

Resumo . . . ix

Abstract . . . xi

List of Tables . . . xv

List of Figures . . . xvii

Acronyms . . . xix

1 Introduction 1

1.1 Objectives and Contributions . . . 2

1.2 Methodology . . . 3

1.3 Thesis Outline . . . 4

2 Concepts and Related Work 5

2.1 Multimorbility . . . 5

2.2 Heart Failure . . . 7

2.2.1 Pathophysiology and Diagnosis . . . 7

2.2.2 Classification . . . 7

2.2.3 Treatment . . . 8

2.3 Electronic Health Records Phenotyping . . . 9

2.3.1 Phenotyping Methods . . . 10

2.4 Clustering methods . . . 13

2.4.1 Clustering Evaluation . . . 14

2.5 Clustering for Disease Phenotyping . . . 16

2.5.1 Clustering Mixed-Type Data in the Electronic Health Record . . . 17

2.5.2 Graphs as a Visualisation Tool . . . 20

2.6 Survival Analysis . . . 21

2.6.1 Kaplan-Meier . . . 21

2.6.2 Cox Proportional Hazards . . . 22

2.7 Summary . . . 23

3 Patient Clustering Workflow 25

3.1 Data . . . 26

3.2 Preprocessing 27
3.3 Patient Clustering 28
3.3.1 Clustering Evaluation 28
3.4 Statistical Analysis and Characterisation of Obtained Clusters 29
3.4.1 Prescriptions 29
3.4.2 Medical appointments 31
3.4.3 Outcomes 31
3.4.4 Survival Analysis 31
3.5 Summary 32

4 Results and Discussion 35

4.1 Characterisation of the Heart Failure dataset 35
4.1.1 Comparison of Female and Male Patients 39
4.2 Clustering Algorithm and Number of Clusters k 42
4.2.1 Choice of cluster number k 43
4.3 Heart Failure Patient Subgroups 43
4.3.1 Prescriptions 47
4.3.2 Medical Appointments 48
4.3.3 Outcomes 52
4.3.4 Cluster Survival Analysis 53
4.4 Summary 56

5 Conclusions 57

5.1 Limitations 58
5.2 Future Work 59

Bibliography 61

List of Tables

2.1 Inter-group proximity measures used for hierarchical clustering. 14
2.2 Overview of HF clustering studies. 18

3.1 ICD-9 Codes and keywords used to identify HF patients from the HLL database. 26
3.2 Feature summary of the Hospital da Luz Lisboa HF dataset. 26
3.3 Medication groups considered and medications included in each one. 30

4.1 Clustering metrics results for the different clustering algorithms. 43
4.2 Clustering metrics results using Hierarchical Clustering with Gower's Distance. 44
4.3 Characterisation of the four resulting clusters obtained from the Hospital da Luz Lisboa. . 45
4.4 Number of medication and medication groups prevalence per cluster and in the entire dataset. 49
4.5 Characterisations of outcomes related variables per cluster and in the entire dataset. . . . 52
4.6 Odds ratio between each cluster and other clusters for the outcomes mortality, hospitalisation and emergency. 52
4.7 Coefficients for the Cox proportional Hazards Model for the outcomes Hospitalisation and Emergency admission. 55

List of Figures

2.1 Example of an undirected graph andof a directed graph 20

3.1 Overview of the proposed approach. 25

3.2 Missing values in initial dataset. 27

3.3 Formula for odds ratio. 32

4.1 Gender, Age and BMI distribution of the dataset of Heart Failure (HF) patients 35

4.2 Prevalence of all comorbidities in the HF dataset. 36

4.3 Prevalence and graph representation of comorbidities. 37

4.4 Histogram for laboratory variables. 38

4.5 Prevalence and graph representation of medical appointments. 39

4.6 Prevalence and graph representation of comorbidities used for clustering by gender. . . . 40

4.7 Medical appointments by gender. 41

4.8 Graph representation of medical appointments by gender. 42

4.9 Boxplots of outcomes hospitalisation and emergency admissions per year. 42

4.10 Cluster-specific percentages of comorbidities. 46

4.11 Graph representation of comorbidities present in the obtained clusters. 47

4.12 Distribution of medical appointments in each cluster. 50

4.13 Graph representation of medical appointments and transisitons in each cluster 51

4.14 Kaplan-Meier survival curves for the outcomes Hospitalisation and Emergency admission. 54

4.15 Cox regression coefficients for Model3. 55

Acronyms

ACEi Angiotensin-Converting-Enzyme Inhibitors.

AFIB Atrial Fibrillation.

ARB Angiotensin Receptor Blockers.

ARNi Angiotensin Receptor/Neprilysin Inhibitors.

BMI Body Mass Index.

BNP B-type Natriuretic Peptide.

CKD Chronic Kidney Disease.

COPD Chronic Obstructive Pulmonary Disease.

EDA Exploratory Data Analysis.

EF Ejection Fraction.

EHR Electronic Health Record.

FAMD Factor Analysis of Mixed Data.

GRF Glomerular Filtration Rate.

HF Heart Failure.

HFmrEF Heart Failure with Mildly Reduced Ejection Fraction.

HFpEF Heart Failure with Preserved Ejection Fraction.

HFrEF Heart Failure with Reduced Ejection Fraction.

HLL Hospital da Luz Lisboa.

HR Hazard Ratio.

HT Hypertension.

ICD International Classification of Diseases.

ICM Ischaemic Cardiomyopathy.

INESC-ID Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento.

LV Left Ventricle.

LVEF Left Ventricle Ejection Fraction.

MCA Multiple Correspondence Analysis.

MICE Multivariate Imputation by Chained Equations.

ML Machine Learning.

NAFLD Non-alcoholic fatty liver disease.

NLP Natural Language Processing.

NT-proBNP N-terminal-pro-BNP.

OB-GYN Obstetrics-Gynecology.

OCPD Obsessive Compulsive Personality Disorder.

OR Odds Ratio.

PAD Peripheral Arterial Disease.

PCA Principal Component Analysis.

PDN Phenotypic Disease Network.

RDW Red Cell Distribution Width.

SGLT Sodium-Glucose Transport Proteins.

SVM Support Vector Machine.

VD Valvular Disease.

Chapter 1

Introduction

Advances in medicine and improvements in living conditions have been driving a continuous increase in life expectancy in most countries over the last century. However, a higher life expectancy does not translate into higher life quality. As life expectancy increases, the percentage of people suffering from more than one chronic condition is increasing dramatically. According to Navickas et al. (2016), this co-occurrence of two or more chronic conditions, defined as multimorbidity, is estimated to affect around 50 million people in the European Union, making it one of the most challenging problems faced by the health sector at the current time.

This dissertation was developed within the Intelligent Care project, a collaboration between Hospital da Luz Lisboa (HLL), Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID), Instituto de Sistemas de Robótica, Carnegie Mellon University and Priberam, which aims to find artificial intelligence solutions to establish the best treatment methodologies for patients with multimorbidity.

Patients with multimorbidity pose a challenge in many ways. From the economic perspective, Rijken et al. (2013) reported that 70% to 80% of healthcare costs are spent on chronic diseases, and Bähler et al. (2015) concluded that patients with more than one chronic condition are associated with substantially higher health care utilisation and costs. From the practitioner perspective, multimorbidity is linked to more complex medical care which makes treatment decisions more difficult and responses to therapy harder to predict. Finally, patients with multimorbidity are also usually associated with poorer health outcomes, poorer quality of life, faster disease progression and higher risk of death.

Despite all negative impacts, current healthcare systems are not prepared to cope with patients with multimorbidity (Colombo et al., 2016). It is necessary to change healthcare from a disease-oriented approach to patient-centred multidisciplinary care, an approach focusing on the patient as a whole, and taking into account the different diseases and the possible interactions among them.

Thus, there is a need to study and characterise populations of patients with multimorbidity. The identification of the patients' condition and specific characteristics, a process denominated as phenotyping, can bring several advantages. The characterisation of patient cohorts and study of the corresponding phenotypes can provide a better understanding of the most common associations and interaction between diseases, providing helpful insights for predicting treatment response. It can also have an important impact on drug development strategies and clinical trials, as it can provide additional information regarding the target population (Council, 2012).

The Electronic Health Record (EHR) is the standardised tool for capturing patients' medical history and constitutes a key source of information for patient phenotyping. An EHRs contains structured data, such as laboratory results, medications, and diagnoses, along with unstructured data, such as radiology reports, progress notes, discharge summaries, and other clinical narratives. Information from structured data can be easily used in phenotyping algorithms, whereas information from unstructured data needs to be extracted and analysed, which poses a challenge by itself. Nonetheless, to obtain clinically significant patient cohorts it is necessary to work with both types of data.

There are several approaches to phenotyping, using rule-based methods, natural language processing, statistical analysis, and machine learning, along with hybrid systems combining different methods. In recent years, research has been shifting from rule-based methodologies to machine learning methodologies (Shivade et al., 2014). Unsupervised learning algorithms, such as clustering, do not not need an *a priori* classification, which means the clusters obtained are based only on the patterns found in the data.

A relevant application for characterising patient cohorts is related to the identification and phenotyping of patient subgroups within a specific disease or condition. Certain conditions, such as HF, Chronic Obstructive Pulmonary Disease (COPD) and Parkinson's have distinct subtypes (Raherison et al., 2018; Thenganatt and Jankovic, 2014). Proper characterisation of patients within each subtype has an impact on therapeutic response and patient outcomes. Several studies have shown that such conditions can greatly benefit from identifying specific patient cohorts since this information can aid in clinical decision making and assess specific risks for each cohort (Francis et al., 2014).

HF is estimated to affect 64.3 million people worldwide, and in developed countries, its prevalence is generally estimated at 1% to 2% of the general adult population (Groenewegen et al., 2020). Despite some improvements over the last years, HF prognosis remains poor and patients' quality of life remains low. The high complexity of HF syndrome and the high heterogeneity of patients makes disease management extremely challenging and can lead to ineffective treatment in some cases. There is a need for an improved characterisation of HF patients, that can help design clinical trials and contribute to defining therapeutic strategies targeted for their individual characteristics.

1.1 Objectives and Contributions

The main goal of this dissertation was to develop a workflow for the identification of subgroups in a population of HF patients with multimorbidity using data from the EHRs, made available by HLL.

The purpose of the workflow was to identify clinically relevant HF patient subgroups, using clustering methods, and to obtain the phenotypes of these groups, that is, to characterise them along with the fields of demographics, comorbidities and laboratory values, present in the EHRs. By doing so, it was also possible to evaluate if the information present in the EHRs is detailed enough to perform a proper characterisation of the patient subgroups.

A secondary goal was to understand to what degree do the patient subgroups differ from each other, and if there are certain subgroups at a higher risk of hospital admission or that have a higher healthcare resource utilisation. This analysis was performed with the help of Dr. Pedro Sarmento, a HF specialist from HLL.

The main contribution resulting from this research project is the proposed workflow for the identification and phenotyping of clinical subgroups in HF, a complex heterogeneous syndrome, along with a thorough characterisation of the HF population from HLL. The application of the workflow to the HF dataset from HLL generated four patient clusters and their clinical phenotypes. The analyses performed, enabled us to understand the heterogeneity of the HF syndrome and to define phenotypes of the identified patient subgroups.

During this research, an article with the early results of the proposed approach, entitled ”Identifying Subgroups in Heart Failure Patients with Multimorbidity by Clustering Analysis”, was produced and presented in the 12th edition of INForum, a Portuguese conference on Informatics. A second article with the resulting analysis from this dissertation is being prepared for submission to a journal.

1.2 Methodology

The first stage of the work was to understand why multimorbidity is a challenge, its impact on patients with multimorbidity, and how can patient phenotyping provide tools to improve care for these patients. Studying the HF syndrome, its pathophysiology and the most common comorbidities associated with it was also fundamental to understand the complexity of the task and to recognise specific characteristics to consider in the development of this work.

The following step was to conduct a revision of the literature on multimorbidity and phenotyping using the EHRs. This research focused particularly on using clustering algorithms to identify patient groups within a heterogeneous disease. Having the work of Nagamine et al. (2020) on HF phenotyping using unstructured EHRs data as a starting point, a review on other papers that also focused on HF, allowed us to understand the most common data features used for HF clustering and the specificities of the disease. Ideas from several previous publications, using different clustering algorithms and including different data sources and data domains, were taken into consideration and inspired the final workflow that has been proposed.

To develop this work, an initial dataset was generated from a database available from HLL as a part of the Intelligent Care project. The Intelligent Care team performed the extraction of HF patient records from the database. Relevant features for the HF clustering were chosen based on literature review and advisory of the HF specialist. Besides the features used for clustering, data from other domains, including prescriptions, medical appointments and outcomes, such as hospitalisations and emergency admissions, were also collected to obtain a detailed characterisation of patients profiles.

The next stage consisted of the design of the workflow for disease phenotyping. The workflow was implemented in the Python programming language, due to its vast public documentation and libraries available for data analysis and machine learning, namely *scikit-learn*[1] and *pandas*[2]. The development of the workflow involved testing several clustering algorithms and different configurations to assess which combination would output the best partition. The quality of the resulting clusters was assessed with three cluster validity indexes and the resulting clusters were also evaluated qualitatively from a clinical perspective.

The final stage of this project involved comparing the obtained clusters against others found in the literature and validating the clusters from a clinical perspective in collaboration with the HF specialist.

1.3 Thesis Outline

The thesis is structured as follows:

- Chapter 2 introduces background concepts needed to understand the rest of the thesis. First, it provides an overview on multimorbidity and on the HF syndrome, describing the disease, diagnosis, possible classifications and treatment options. Afterwards, this chapter discusses the EHRs phenotyping and patient clustering methods, including a review on approaches related to EHRs phenotyping, with a special emphasis on HF studies.

- Chapter 3 describes the proposed workflow for the identification of HF patient groups and subsequent phenotyping. Starting with the details of the structure of the data extracted from the HLL database, and the preprocessing stage, it follows to an initial exploratory analysis of the dataset. The different clustering evaluation metrics used are also presented. Then, it describes the proposed clustering method and the characterisation and statistical analysis of the resulting clusters.

- Chapter 4 presents the results from the patient clustering workflow. It starts with a brief characterisation of the dataset, including a comparison between female and male patients. It also presents the evaluation metrics results and discusses the choice of the clustering algorithm and the number of clusters. Finally, the chapter gives a detailed characterisation of the clusters obtained in all domains analysed, tracing a profile for each cluster and discussing the results obtained from a clinical perspective.

- Chapter 5 summarises the main conclusions of this work and its limitations, and presents possibilities for further developments.

[1] https://scikit-learn.org/stable/

[2] https://pandas.pydata.org/

Chapter 2

Concepts and Related Work

The care of patients with multimorbidity represents a complex issue in the current healthcare system. Phenotyping using Electronic Health Record (EHR)s, also known as electronic phenotyping, can provide relevant insights into patients characteristics, needs and risks, with the end goal of improving patient care. This chapter describes relevant concepts on electronic phenotyping with a particular focus on patients with Heart Failure (HF). Section 2.1 provides an overview on multimorbidity, how it affects patients, healthcare institutions and healthcare providers, and how future work should be oriented to improve care for these patients. Section 2.2 describes the HF syndrome, the current classification and treatments. Section 2.3 presents the concepts needed to understand electronic phenotyping and explains the different methodologies developed in this field. Section 2.4 briefly introduces clustering methods and clustering evaluation, and Section 2.5 reviews previous work in the field of clustering for patient phenotyping, with a focus on HF. Section 2.6 addresses survival analysis and how it can be used to compare different patient groups.

2.1 Multimorbility

Multimorbidity, defined by the World Health Organization (WHO), as the ”coexistence of two or more chronic conditions in the same individual” is one of the most challenging problems of public health at the present time (WHO, 2016). Multimorbidity is estimated to affect around 66.2% of Europe's population aged over 50 years old, and, with the increase in life expectancy and the ageing of populations, this percentage is expected to rise in all populations of patients around the world. According to Navickas et al. (2016), patients with multimorbidity generate the highest expenses while experiencing the least benefits from healthcare systems.

One of the aspects of multimorbidity that is particularly important to study is the risk factors. It is necessary to understand which populations have a higher predisposition and which type of socioeconomic factors increase the likelihood of developing multimorbidity in order to develop preventive strategies. As reported in Navickas et al. (2016) and in most studies concerning multimorbidity, ageing is the most preponderant risk factor, however, multimorbidity is not restrained to the older population. The prevalence of multimorbidity was reported to increase with the ageing of the population but also to be present in all age groups, starting at 10% in the 0-19 years group. WHO (2016) also reveals the existence of a relationship between underprivileged individuals with lower socioeconomic status and low education level and the presence of multimorbidity, results that have been replicated by Broeiro-Gonçalves (2015). Other risk factors identified by the previous studies include smoking, physical inactivity, high body mass index (BMI), and hypertension.

Patients with multimorbidity represent a high burden to healthcare systems, both in terms of utilisation and cost. A study performed by Bähler et al. (2015) on a Swiss population of 229 493 individuals aged 65 or older reported that the total healthcare costs were 5.5 times higher in patients with multimorbidity. Additionally, each chronic condition was associated with increased costs of 33% and a 3.2 increase in the number of consultations. According to Rijken et al. (2013), 70-80% of health care costs are spent on chronic diseases, which corresponds to €700 billion in the European Union. This is strongly related to the current organisation of healthcare systems. The majority of existing care delivery models are disease-specific or organised around acute episodes, and therefore are unsuited to the needs of the increasing number of people with multiple health problems.

On an individual level, multimorbidity impacts patients life quality, well-being and physical and mental health. Despite therapeutic advances, these patients may be at greater risk of complications and represent a greater difficulty in predicting the response to therapeutic regimes. Patients with multimorbidity have more frequent and complex interactions with health care services. As pointed out by Colombo et al. (2016), these patients had a much higher average number of consultations when compared to patients without multimorbidity (15.7 and 4.4, respectively) and were 5.6 times more likely to be hospitalised. According to WHO (2016), multimorbidity is also associated with faster disease progression and higher risk of safety issues, due to patient complexity and to the administration of many drugs, which may lead to poor medication adherence or adverse drug events.

Although data about multimorbidity and the most common combination of diseases, multimorbidity clusters, is still limited, it is known that multimorbidity patterns are highly heterogeneous and that there are a number of different combinations of conditions that patients can experience. There are cases where co-existing chronic diseases might be similar in their aetiology or have a similar approach to treatment. One example of this is coronary heart disease and cerebrovascular disease, which share a common aetiology, high blood pressure. In other cases, the conditions appear unrelated to one another or require separate management strategies. The study of multimorbidity clusters and respective characterisation can help to understand the most common association between diseases and the influence that some conditions might have on others and in this way it can bring great value to treatment choice, clinical trials design and drug development.

In terms of future research in the field of multimorbidity, the report developed by The Academy of Medical Sciences (2018) summarises the available evidence on multimorbidity and highlights key evidence gaps, recommending a series of research priorities. The recommendations include: trends and patterns in multimorbidity; which multimorbidity clusters cause the greatest burden; determinants of the most common clusters of conditions and strategies for prevention of chronic conditions that contribute to the most common multimorbidity clusters. These recommendations highlight the importance of understanding the co-existence of chronic disease to develop more effective prevention and treatment strategies in addition to improving patients life quality and well being overall. Alongside with research, it is also necessary to rethink healthcare systems and shift towards a framework of patient-centred, integrative medicine.

2.2 Heart Failure

HF is a complex clinical syndrome characterised by the inability of the heart to maintain a cardiac output that suffices the body demands of oxygen and blood. According to Groenewegen et al. (2020), HF is estimated to affect more than 64.3 million people worldwide and Savarese and Lund (2017) report it to be responsible for around $31 billion in health expenditures.

2.2.1 Pathophysiology and Diagnosis

HF results from structural and/or functional cardiac abnormality most commonly in the myocardium, but also in the pericardium, endocardium, heart valves, coronary arteries, or abnormalities of heart rhythm and conduction, that cause impairment of either ventricular filling or blood ejection. In most HF patients, the symptoms arise from cardiopulmonary or systemic congestion and aggravate with the elevation of natriuretic peptides. HF has different etiologies and pathophysiology rather than a specific disease. The most common are ischaemic heart disease, hypertension, and diabetes. As the condition progresses, patients can suffer from a number of symptoms and signs including fatigue, breathlessness, reduced exercise tolerance, orthopnea, ankle swelling, paroxysmal nocturnal dyspnoea, a decline of cognitive function and cold extremities.

As described by The American College of Cardiology Foundation/American Heart Association in Yancy et al. (2013), the diagnosis of HF is not straightforward as there is no single diagnostic test. It is predominantly a clinical diagnosis based on patient history and physical examination. Patients often have a history of arterial hypertension, ischaemic heart disease, diabetes mellitus, alcohol abuse or cardiotoxic chemotherapy. The recommended diagnostic tests consist of a chest X-ray, echocardiography, and plasma levels of B-type Natriuretic Peptide (BNP) or N-terminal-pro-BNP (NT-proBNP). BNP is a hormone released by the ventricles whenever the heart undergoes stress, whether chronic or acute, in an attempt to compensate the vasoconstrictor systems that are activated in these situations. It has great prognostic value in the context of HF and is a powerful rule-out test. Echo-cardiogram, for its availability, is considered the key investigation for the assessment of cardiac dysfunction.

2.2.2 Classification

In the new Universal definition of HF (Bozkurt et al., 2021), the condition is classified in four phenotypes based on the Ejection Fraction (EF): *i)* Heart Failure with Reduced Ejection Fraction (HFrEF), when Left Ventricle Ejection Fraction (LVEF) $<$ 40%, *ii)* Heart Failure with Mildly Reduced Ejection Fraction (HFmrEF), when LVEF is 41-49%, *iii)* Heart Failure with Preserved Ejection Fraction (HFpEF), when LVEF $>$ 50% and *iv)* HF with recovered EF, when LVEF $>$ 40%.

EF is a measurement, usually obtained using an echocardiogram, of the percentage of blood that the heart pumps out with each contraction. A healthy individual usually has a preserved EF, normally above 70%.

In HFrEF (also called systolic HF) the EF the Left Ventricle (LV) contracts poorly and empties inadequately, leading to increased diastolic volume and pressure and reduced ejection fraction. Predominant systolic dysfunction is common in HF due to myocardial infarction, myocarditis, and dilated cardiomyopathy.

In HFpEF (also known as diastolic HF), there is an increased diastolic ventricular stiffness, which slows LV relaxation, increases LV diastolic filling pressures and limits cardiac output. However, global contractility of the heart remains normal which explains the preserved ejection fraction. HFpEF has been studied in more depth in the past years it has been shown to be a complex, heterogeneous, multi-organ systemic syndrome, often with multiple co-occurring pathophysiologies. Current data suggest that multiple comorbidities (e.g. obesity, hypertension, diabetes, chronic kidney disease) lead to systemic inflammation, generalised endothelial dysfunction, cardiac microvascular dysfunction and ultimately molecular changes in the heart that cause increased myocardial fibrosis and ventricular stiffening.

2.2.3 Treatment

Treatment of HF has the main goal of improving prognosis, reducing mortality and relieving the patient symptoms, reducing morbidity. However, according to Kemp and Conte (2012), despite existing treatments, the mortality rate of HF remains very high. About 50% of patients die within five years of diagnosis of the disease, 78% have at least two hospital admissions per year and patients take on average six HF-related medications, which leads to an annual cost of $10–38 billion.

The treatment includes lifestyle modifications as well as medical therapies. Patients are encouraged to lose excess weight, reduce the consumption of sodium, refrain from tobacco and alcohol use, and engage in any tolerated physical activity. The prescribed medical therapies are highly dependent on the patient's disease classification (HFpEF or HFrEF), progression and severity, and might include pharmacological treatment and surgical procedures. Medications used to the relieve of symptoms are diuretics, nitrates or digoxin and medications used in the prolonged treatment and improved survival are Angiotensin-Converting-Enzyme Inhibitors (ACEi)s, beta-blockers, aldosterone antagonists, Angiotensin Receptor Blockers (ARB)s or Angiotensin Receptor/Neprilysin Inhibitors (ARNi)s. Examples of device and surgical procedures performed in HF patients are cardiac resynchronization therapy (CRT), coronary revascularization, Surgical Ventricular Remodelling (SVR), Ventricular Assist Device (VAD) implantation, and, in extreme cases, heart transplantation.

Regarding pharmacological treatment response and outcomes, Inamdar and Inamdar (2016) reported that patients with reduced ejection fraction (HFrEF) respond well to the standardised treatment and have a higher probability of a better prognosis. On the other hand, few options for pharmacological treatment have been shown to be effective for patients with preserved ejection fraction (HFpEF). Indeed, they do not respond well to all the same drugs used for those with reduced ejection fraction. The difference in response to treatment from patients with HFpEFand HFrEF is related to the pathophysiology of each of these groups explained previously.

Numerous pharmacological clinical trials have tried to discover a treatment that would improve morbidity and mortality in patients with HFpEF, but so far only one study, the EMPEROR-preserved study with empagliflozin, showed significant results (Anker et al., 2020). A reason for this might be related to the design of the clinical trials and to the high heterogeneity in the patient population, with differences in HF etiologies and stages of the disease.

HF is a complex syndrome with a high prevalence of multiple chronic conditions and represents a heavy burden to healthcare systems. It is necessary to further study and characterise subgroups of patients within the HF cohort to improve treatment options and outcomes.

2.3 Electronic Health Records Phenotyping

In the past years, EHRs have been adopted by a vast number of countries which has generated a large amount of clinical data. The EHR is a standardised information technology tool that contains a patient's medical history, diagnoses, medications, treatment plans, laboratory and test results. Data in the EHRs can be either structured, as diagnosis codes and laboratory results, or unstructured, such as exams reports and clinical notes. One of the most promising applications for EHRs is to identify groups of patients with certain characteristics, a process named electronic phenotyping.

The definition of electronic phenotyping is very broad, and depending on the study it can correspond to different concepts. In some studies, electronic phenotyping is simply identifying a specific disease in patients, whereas in others, phenotyping is used to distinguish patient subgroups within a single disease, by providing the most detailed characterisation of the patient. However, definitions always include the characterisation of clinical features in the patient, which can include demographics, as age and sex; the presence of a certain disease or condition; exposures, as medication, smoking, and alcohol use; and outcomes, such as mortality and hospitalisation. In this section, we will provide examples for the phenotyping of a specific disease and the phenotyping within a single disease.

There are several areas in which electronic phenotyping can be of great use. Banda et al. (2018) performed a literature review and identified the main methods and applications of patient phenotyping using EHRs data. They concluded that electronic phenotyping was being applied in cross-sectional studies, e.g., for hospital resource allocation, adherence to diagnostics and treatment guidelines; association (case-control/cohort) studies, e.g for clinical effectiveness research and predictive modelling; and for experimental studies, e.g for clinical trial recruitment and adaptive/randomised trials. Moreover, the characterisation of patient cohorts and study of the corresponding phenotypes can provide a better understanding of the most common associations and interactions between diseases, providing helpful insights for predicting clinical outcomes. Shivade et al. (2014) pointed out that better characterised cohorts can also have an important impact in drug development strategies and clinical trials, with additional information regarding the target population.

However, due to the heterogeneity of the data, frequent incompleteness and dynamic nature of EHRs, the identification of phenotypes using EHRs is still a complex task. This complexity also contributes to the difficulty in creating a standardised tool for electronic phenotyping that could be used in different institutions. Other barriers to such a tool are the variability in how data are inserted into the EHR in each clinical site, the lack of data quality assessment measures and administrative roadblocks (Weiskopf and Weng, 2012). Consequently, each institution develops their own methods, adapted to their specific requirements, and comparison of phenotyping results across different institutions becomes increasingly difficult.

Electronic phenotyping methods can use structured data, unstructured data, or a mix of both. The majority of methods uses structured data only as, for example, codes from the International Classification of Diseases (ICD) System. Methods that use unstructured data, such as notes from exams or appointments, are able to extract more detailed and accurate information. However, in order to obtain this information, it is necessary to overcome the challenges of processing clinical text, which is written using very specific terminology, uses abbreviations and often contains misspellings. Tayefi et al. (2021) reviewed the possibility of using unstructured data in addition to structured data and concluded that even though there are still unsolved challenges when dealing with unstructured data, there are many opportunities to develop methods that incorporate both types of data to extract helpful information and provide better diagnosis and decision support tools.

The ICD system provides a standard tool for reporting diseases and health conditions internationally. It is a comprehensive, hierarchical listing of diseases, disorders, injuries, and other associated health problems. According to WHO (2018), it allows for easy storage, retrieval and analysis of health information for evidence-based decision-making, to compare health information between different healthcare institutions and countries, and also to compare information across different time periods. Due to its widespread use and frequent update, the ICD system has become suitable for a variety of uses in health, as for example, causes of death, external causes of illness, medications, primary care and family medicine and rare diseases.

2.3.1 Phenotyping Methods

Over the years different approaches have been developed to identify patient cohorts using data from EHRs. The earliest methods consisted of rules applied to structured data, known as rule-based methods. From there, more complex approaches have been developed, making use of emerging technologies, such as Machine Learning (ML) and Natural Language Processing (NLP), and integrating clinical information from unstructured data. Shivade et al. (2014) described phenotyping systems and techniques based on EHR data and identified three main approaches: Rule-Based, NLP and ML.

Rule-based systems

Rule-based systems are a set of logical constraints and rules that work as inclusion or exclusion criteria. The algorithms used in rule-based phenotyping range from simple pattern matching to more complex techniques, which include multiple logical steps and combine different operations, such as boolean or comparative. These systems can be built according to clinical judgement or healthcare guidelines, and use different sources of data. The most common ones are diagnosis codes and patient characteristics, but some systems also make use of billing codes, number of visits or hospital admissions, laboratory results, and prescriptions. The rules used in the algorithm are usually defined based on clinical expertise or in guidelines and recommendations of health recommendations.

When it comes to phenotypes belonging to a specific subdomain, that have clear diagnosis and procedure codes, rule-based methods perform well. Esteban et al. (2017) identified cardiovascular and cerebrovascular disease cases using only clinically relevant terms from the International Classification of Primary Care, Second Edition (ICPC-2) and the ICD-10. These terms included not only diagnosis codes but also signs, symptoms, and procedures, as queries that combine different structured data fields tend to show higher performance than queries using a single code search. Wei et al. (2015) observed that using multiple disease codes together with medication data improved precision and query performance compared to using a single diagnosis code for a range of different diseases.

Advantages of rule-based methods, as pointed out by Alzoubi et al. (2019), are the simplicity of the construct, the accuracy when using small datasets, and the interpretability of the algorithm. However, the process requires profound clinical knowledge, and most times it is only used in a specific dataset, lacking validation and thus the possibility of generalisation for other datasets. Moreover, these methods are usually limited to simpler phenotypes, as higher complexity causes a decrease in phenotype performance, and is dependent on the structure of the datasets.

Natural Language Processing methods

The majority of information present in EHRs is in the form of unstructured data, which includes clinical notes, reports from examinations or discharge summaries. This information contains relevant phenotypic characteristics, that are often not represented in structured data fields. Automated solutions using NLP allow performing tasks such as summarisation, entity recognition, and relationship extraction. NLP methods have been effectively applied to a range of problems in different areas. However, the specificity, poor phrase structure, misspellings, ambiguity and other aspects of the clinical text in EHR pose an additional challenge (Pathak et al. (2013)).

The first methods developed to identify concepts from clinical text consisted of pattern-matching against standard terminologies. In the past years, clinical NLP methods went through several developments and are now focusing on analysing semantic relationships within the text, allowing for better identification of phenotypic characteristics. Once the relevant terms are extracted from free text fields, patients are classified into cohorts using either a rule-based or a ML model.

Kreimeyer et al. (2017) found that for simpler tasks as, for example, finding an expression pattern, rule-based methods had acceptable performance, with the downside that the rules had to be specific for the institution developing them, not having the possibility to be scaled. The study also reported a growth in ML approaches as more public datasets are becoming available. Carroll et al. (2011) created a Support Vector Machine (SVM) classifier for rheumatoid arthritis combining concepts and medications extracted using NLP with structured data.

Machine Learning approaches

Due to the growing amount of data, the applications of ML in healthcare have been rising in a wide number of domains. In the field of electronic phenotyping, ML methods are used due to their scalability and ability to infer patterns from the data, reducing the effort needed from clinical experts. These approaches can work with both structured and unstructured data, and use supervised, semi-supervised, or unsupervised learning algorithms. Supervised learning uses a ground truth, which means that the output values for the samples are known *a priori*. As a result, the purpose of supervised learning is to find a function that best approximates the relationship between input and output observable in the data given a sample of data and desired outcomes. This strategy, however, is dependent on annotated materials, which are expensive, complex, and time-consuming to provide in the medical field. Unsupervised learning, on the other hand, uses unlabeled data and seeks to understand the inherent structure existing in a set of data points. The case of semi-supervised learning refers to learning problems where the model learns from only a small number of labelled examples and can classify a large amount of unlabeled data.

An application of supervised learning can be seen in Pestian et al. (2016), which compared SVM and Naives Bayes algorithm in an identification task for pediatric epilepsy, with SVM achieving the best performance. Other supervised models commonly used include Bayesian Networks, logistic regression, decision trees and artificial neural networks (Banda et al. (2018)).

Unsupervised learning algorithms are able to process large volumes of data without needing manual labelling. However, the lack of labelling means that there is no ground truth for the phenotypes, and thus the validation of phenotypic groups obtained with this approach remains quite challenging. Unsupervised learning also allows detecting hidden patterns within the data which can lead to the discovery of new phenotypes. One technique that is frequently used when applying unsupervised learning algorithms to electronic phenotyping is tensor factorising of EHRs. Ho et al. (2014) developed a model using non-negative tensor factorisation to infer phenotypes from the interaction of diagnosis and medications in patients. The study found that the phenotypes generated by the model had robustness, stability and were clinically meaningful. Another unsupervised technique commonly used is clustering, used by Bleecker et al. (2014) to identify subgroups of asthma patients using clinical data.

In conclusion, there is a clear need to develop methods for electronic phenotyping that can make use of structured and unstructured data found in EHRs. NLP techniques are especially important to take advantage of information stored in unstructured data and bring great value when combined with rule-based or ML methods. Rule-based methods are still dominant in most institutions since most analyses are performed in specific patient populations. In order to develop portable solutions the inclusion of ML methods is necessary.

2.4 Clustering methods

Clustering is one of the ML algorithms used in electronic phenotyping. One advantage of using clustering algorithms is that it is possible to integrate both structured and unstructured data from EHRs. To use unstructured data, clustering algorithms can be combined with NLP algorithms. NLP algorithms can be used to retrieve important information from clinical text, such as diseases or symptoms, that can then be used in the clustering algorithm.

Clustering is an unsupervised machine learning technique used to discover natural occurring groups of a set of objects, so that objects within the same group/cluster are very similar, and elements in different groups are distinct from each other. However, there is no formal definition for the term cluster. In practice, a cluster is a subjective structure whose relevance and interpretation are dependent on domain knowledge. Gordon (1999) defines the notion of a cluster in terms of internal cohesion (homogeneity) and external isolation (separation). To define similarity between the elements, a distance function is used, which needs to be defined considering the context of the problem at hand. We can divide cluster analysis into two main types of methods according to the way groups are formed: hierarchical and partitional.

Partitional Clustering

Partitional clustering algorithms are characterised by the need to define an initial partition and by their flexibility, since the elements can be changed from one group to another during the execution of the algorithm. These algorithms work in an iterative way and relocate data points between clusters until an optimal partition is reached. The optimal partition is attained by optimising certain clustering criteria. Some of the most commonly used partitional clustering algorithms are K-Means, PAM (K-Medoid) and CLARANS (Clustering Large Applications). Partitional algorithms are efficient but sensitive to initial conditions and outliers.

K-means is one of the simplest and most widely used clustering methods. It divides the data into multiple groups, with data points from the same cluster being as similar as possible (high intra-class similarity), and data points from different clusters being as dissimilar as possible. The objective is to minimise the sum of squared error over all k clusters, given by

$$J(C) = \sum_{k=1}^{K} \cdot \sum_{x_i \epsilon c_k} \|x_i - \mu_k\| \tag{2.1}$$

where μ_k is the centroid of cluster c_k. The centroid, or cluster centre, is defined as the arithmetic mean of all points belonging to the cluster and each point must be closer to its own centroid than to any other. In the first iteration, a predefined number of k centroids are initialized randomly. In each iteration, every data point is allocated to the nearest cluster and the centroids are recomputed using the new data points in the cluster. As referred by Jain (2010), the minimisation of the equation 2.1 for any given k is an NP-hard problem and K-means can only converge to local minima. Consequently, different initialisations can result in different final clusters. One approach to deal with this is to run the algorithm multiple times with different initialisations and choose the one that outputs the partition with the smallest sum of squared errors.

Several extensions or different versions of k-means have been developed over the years. The metric used in k-means to compute the distance between data points and cluster centroids is usually the Euclidean metric. Other metrics have been proposed, such as the L1 norm, used in Kashima et al. (2009) for speech processing or the Bregman distances used in Banerjee et al. (2007) to cluster high-dimensional vectors. Fuzzy c-means is an adaptation of k-means where each data point is not restrained to belong to a single cluster and instead can be part of multiple clusters, having a membership score for each one (the higher the score the higher the membership to that specific cluster).

Hierarchical Clustering

In hierarchical clustering the data partitioning is performed in a sequential manner instead of in a single step. It requires a series of fusions or partitions and can be divided into agglomerative methods, in the case of fusions, or divisive methods, in the case of partitions. In Agglomerative Hierarchical Clustering (AHC), each instance starts in its own isolated cluster, and instances are then merged with similar instances to form similarity clusters until only one cluster with all instances is formed. In Divisive Hierarchical Clustering (DHC) all instances start in one cluster and are successively divided into smaller groups.

The most commonly used method in hierarchical clustering is AHC. At each step of AHC, clusters or groups are fused with the ones that are closest or most similar according to a predefined distance/similarity measure. There are many different possibilities of distance or similarity definitions between two cluster groups, the most common distance measures are represented in Table 2.1.

Table 2.1: Inter-group proximity measures used for hierarchical clustering. Adapted from Everitt et al. (2011).

Method	**Distance between clusters defined as**	**Formula**
Single linkage	Minimum distance between pair of objects, one in one cluster, one in the other	$D(c_1, c_2) = min_{x_1 \epsilon C_1, x_2 \epsilon C_2} D(x_1, x_2)$
Complete linkage	Maximum distance between pair of objects, one in one cluster, one in the other	$D(c_1, c_2) = max_{x_1 \epsilon C_1, x_2 \epsilon C_2} D(x_1, x_2)$
Average linkage	Average distance between pair of objects, one in one cluster, one in the other	$D(c_1, c_2) = \frac{1}{\lvert c_1 \rvert} \frac{1}{\lvert c_2 \rvert} \sum_{x_1 \epsilon c_1} \sum_{x_2 \epsilon c_2} D(x_1, x_2)$
Ward’s criterion	Increase in sum of squares within clusters, after fusion, summed over all variables	$TD(c_1 \cup c_2) = \sum_{x_1 \epsilon c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$

2.4.1 Clustering Evaluation

Since clustering is simply a partition of data, when a clustering algorithm is applied it usually finds clusters in the data, regardless if there are any clusters present or not. Moreover, even if there are naturally occurring clusters in the data, the user must specify the number of clusters, k, that the algorithm must find. Cluster evaluation or cluster validity refers to methods that evaluate the results of cluster analysis in a quantitative way, allowing to assess the quality of the clusters and helping the user to understand what number of clusters k to use.

According to Jain (2010), there are three different criteria to define cluster analysis indices: internal, external, and relative. Internal criteria indices evaluate if the structure imposed by the clustering is suitable for the data and use only the data being clustered. External criteria indices require information regarding the true class or ground truth labels. They assess the clustering by comparing the true class labels with the labels obtained through clustering. However, there are many situations where the ground truth is not available. Finally, relative criteria indices measure the algorithm performance comparing different partitions obtained with different algorithms or with different parameters for the same algorithm.

Several internal indices have been proposed to help the user choose the number of clusters k and which clustering algorithm to use. These measures evaluate the performance of a clustering algorithm without any external information. The majority involves running the clustering algorithm with different values of k and assessing the best value based on a predefined criterion. Different criteria may output different values for k, as these criteria might value different components, for example, some might value more the intracluster cohesiveness while others give more weight to the separation inter clusters. Arbelaitz et al. (2013) reviews thirty of the most used internal clustering validation measures such as Silhouette score, Calinski-Harabasz index, Dunn's index, and Davies-Bouldin score, and validates their performance in different application scenarios. Below a more detailed description of Silhouette score, Calinski-Harabasz index, and Davies-Bouldin is provided, as these indexes were in the top scoring indexes of the evaluation by Arbelaitz et al. and will be used for clustering evaluation in this dissertation.

The silhouette score, by Rousseeuw (1987), is calculated using the mean intracluster distance and the mean nearest-cluster distance for each instance, the higher the value, the higher the performance. It is given by:

$$Silhouette\ score(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.2}$$

where

a(i) = average intracluster distance, i.e the average distance between a given point i within a cluster.
b(i) = average intercluster distance,

The second index used, by Caliński and Harabasz (1974), is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion, as for the silhouette score, a higher value translates into higher performance. For a set of data E of size n_E, clustered in k clusters, the Calinski-Harabasz score is given by:

$$CH = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n_E - k}{k - 1}\ k = 2, .., k_{max} \tag{2.3}$$

with

$$W_k = \sum_{q=1}^{k} \sum_{x \epsilon X_q} n_q (x - \bar{x}_q)(x - \bar{x}_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q(\bar{x}_q - \bar{x})(\bar{x}_q - \bar{x})^T$$

where $tr(B_k)$ is trace of the between group dispersion matriz and $tr(W_k)$ is the trace of the within-cluster dispersion matrix. C_q is the set of points in cluster 1, c_q the center of that cluster, c_q the center of E (data) and n_q the number of instances in cluster q.

Finally, the Davies-Bouldin score, measures the average similarity between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves (Davies and Bouldin, 1979). Values closer to zero indicate a better partition. Considering a set of data E of size n_E, clustered in k clusters, the Davies-Bouldin score can be computed as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}(\frac{S_i + S_j}{M_{i,j}}) \tag{2.4}$$

with

$$S_i = (\frac{1}{T_i}\sum_{j=1}^{T_i} |x_j - A_i|^p)^{\frac{1}{p}}$$

$$M_{i,j} = \|A_i - A_j\|_p = (\sum_{k=1}^{n} |a_{k,i} - a_{k,j}|^p)^{\frac{1}{p}}$$

where S_i is a measure of scatter within the cluster and $M_{i,j}$ is a measure of separation between cluster C_i and Cluster C_j. X_i is an instance of cluster C_i, A_i is the centroid of cluster C_i, T_i is the size of the Cluster C_i and $a_{k,i}$ is the k^{th} element of A_i. The value of p is usually 2, making the Davies-Bouldin Index defined as a function of the Euclidean distance.

2.5 Clustering for Disease Phenotyping

Defining and characterising chronic diseases patient' groups can be an important aid to understanding the most common associations and implications between diseases, helping to provide better care to patients with multimorbidity. In the clinical field, clustering has been used to identify patient groups with similar characteristics, in a general population or within a single disease.

Clustering has been widely applied to populations with multimorbidity, in an attempt to find disease groups that co-occur more frequently.

Vetrano et al. (2020) studied a population of 2931 patients during a course of 12 years, identifying multimorbidity clusters, detecting clinical trajectories and tracing the clusters' evolution. The study found five clinically meaningful clusters and one unspecified cluster. It was observed that throughout the years the patients in the unspecified cluster would move to one of the other clusters as other diseases developed. They concluded that multimorbidity clusters can help to identify patient groups with similar prognosis and treatment needs and thus assist healthcare professionals in designing appropriate treatment plans and targeting preventive strategies.

In another study by Violan et al. (2018), around 400 thousand patients with multimorbidity were divided by sex and clustered using K-means on EHRs data. The clustering resulted in six multimorbidity patterns for each gender, with the most prevalent pattern including coincident diseases for both male and female patients. The phenotypes defined by the clustering analysis were consistent with clinical practice and presented similarities to previous studies.

As for discovering phenotypes within a single disease, several studies have used clustering methods to identify clinically relevant patient subgroups in complex conditions such as HF, Chronic Obstructive Pulmonary Disease (COPD), Dementia and Parkinson's Disease, leading to important insights regarding disease pathophysiology.

Vandromme et al. (2020) identified 13,290 patients with Non-alcoholic fatty liver disease (NAFLD) and gathered data from multiple sources, including demographics, diagnosis, procedures, laboratory tests, medications, and vital signs. They applied hierarchical clustering using Ward's criterion for minimising the increase in variance during cluster merging. The study identified five patient subgroups, and using descriptive statistics and survival analysis it was possible to observe that the subgroups were clinically different and associated with different rates for several outcomes considered. The authors state that the novel disease subtypes identified can be used to risk-stratify patients and aid in treatment management.

Using a different approach, Raherison et al. (2018) investigated the relationship between COPD severity and the presence of other comorbidities. They applied Principal Component Analysis (PCA) and non-hierarchical clustering to clinical data from 584 patients and identified five clusters with a different prevalence of the comorbidities studied. The study reported that the presence of comorbidities contributed to disease severity and had an impact on how the disease was managed.

There have also been numerous studies focusing on HF. Ahmad et al. (2014) applied Ward's hierarchical clustering to 2,331 HF patients and identified four clusters whose patients characteristics varied greatly in measures of demographics, symptoms, comorbidities, HF aetiology, quality of life, among others. Using Cox proportional hazards modelling they also found differential associations for hospitalisation and mortality risks between clusters. Gulea et al. (2021) used Latent Class Analysis to cluster HF patients based only on their comorbidities and obtained five clusters that exhibited differences in the risks of hospital admission, mortality, and healthcare resource utilisation.

Given the focus of this dissertation on HF patients with multimorbidity, Table 2.2 summarises the above referenced studies along with a few others that have been proposed to address HF phenotyping, focusing on the ones that used clustering methods.

2.5.1 Clustering Mixed-Type Data in the Electronic Health Record

As seen in Section 2.3, EHRs can have very heterogeneous data. Consequentially, the data extracted from EHRs can contain a mixture of categorical and continuous data. Clustering mixed data is challenging since it is impossible to directly apply mathematical operations on the values of these datasets, such as summing or average (Ahmad and Khan, 2019).

Table 2.2: Overview of HF clustering studies.

Author	Data	Methods	Results
Nagamine et al. (2020)	Clinical notes found in each patient EHR (e.g., diseases, signs, symptoms, conditions)	Frequency-inverse document frequency; K-means clustering with cluster bootstrapping to determine best value of k	Clinically interpretable hierarchy of subgroups characterized by similar HF manifestation.
Shah et al. (2014)	Demographics, Age, Physical characteristics, Laboratory, Echocardiography, Electrocardiography	Hierarchical clustering using Euclidean distance and average linkage score; Penalized model-based clustering; Cox regression	3 distinct groups that differed markedly in clinical characteristics, cardiac structure/function, invasive hemodynamics, and outcomes
Ahmad et al. (2014)	Demographics, Medical history, Laboratory, QOL scores, exercise capabilities	Wards minimum variance method; Cox regression	4 clusters that varied considerably in the clinical variables studied, differential associations were observed for hospitalisation and mortality risks between clusters

Gulea et al. (2021)	Comorbidities	Latent class analysis;
		Kaplan-Meier curves;
		Cox regression;
		Negative binomial analysis
			5 comorbidity clusters that exhibited differences in the risks of hospital admission, mortality, and healthcare resource utilisation

Ahmad et al. (2018)	Age, Creatinine, Hemoglobin, Weight, Heart Rate, Systolic Blood Pressure, Mean Arterial Pressure, Income	K-means clustering with silhouette score to determine best value of k	4 clinically relevant subgroups of HF with marked differences in 1-year survival

Foss et al. (2019) reviewed the advantages and disadvantages of some of the methods identified in the literature to deal with mixed-type data using theoretical and empirical approaches. One of the mentioned strategies was using hybrid distance approaches, that is, using specific distance functions designed for mixed-type data. Other approaches include performing data transformation methods, such as discretisation or dimensionality reduction, and using statistical mixture models. The study concluded that no clustering approach stood out from the others as the efficiency of the algorithm is highly case dependent and so the decision has to be done case to case.

Concerning the use of specific distance functions, Singh et al. (2018) performed multimorbidity clustering and showed that computing Gower's dissimilarity prior to clustering is helpful when working with mixed-type data. Gower's distance is a dissimilarity that measures the similarity of two items with mixed numeric and non-numeric data (Gower, 1971). The Gower distance for two instances x and y is given by:

$$d_G(x,y) = \frac{\sum_{j=1}^{m} w_j \cdot f_j(x_j, y_j)}{\sum_{j=1}^{m} w_j} \tag{2.5}$$

with

$$f_j(x_j, y_j) = \begin{cases} \frac{|x_j - y_j|}{r_j} & \text{, if } x_j \text{ and } y_j \text{ are interval scales} \\ I\{x_j \neq y_j\} & \text{, if } x_j \text{ and } y_j \text{ are categorical scales} \end{cases}$$

where m is the number of features, w_j is the feature weight and I is the indicator function, that is, I is 1 if x_j and y_j are equal and 0 otherwise. The weights w_j were considered 1 for all features.

Dimensionality reduction refers to techniques that have the objective of reducing the number of attributes in a dataset while preserving as much of the variation in the original dataset as possible (Reddy et al. (2020)). There are two types of dimensionality reduction approaches, the first focuses on feature selection, using scoring or statistical methods to select which features to keep and which features to delete, and the second focuses on feature combination, applying a transformation to combine features from the original dataset.

PCA is a linear dimensionality reduction algorithm, belonging to the feature combination methods. The algorithm uses linear combination to transform a set of correlated variables into a smaller number of uncorrelated variables, called principal components (Jolliffe, 2011). PCA is usually applied to numerical data. Multiple Correspondence Analysis (MCA) is a dimensionality reduction algorithm suitable for categorical data, that represents data as points in a low-dimensional Euclidean space (Abdi and Valentin, 2007). There are also dimensionality reduction algorithms specific to deal with mixed-type data, as is the case of Factor Analysis of Mixed Data (FAMD). The FAMD algorithm can be seen as a mix between PCA and MCA, it acts as PCA quantitative variables and as MCA for qualitative variables (Husson et al., 2008).

Verdonschot et al. (2020) who studied Dilated Cardiomyopathy patients phenotypes and used PCA prior to hierarchical clustering on a dataset, provide one example of applying dimensionality reduction prior to clustering.

2.5.2 Graphs as a Visualisation Tool

When clustering data with a high number of features, it becomes hard to visualise the resulting clusters and to understand if they present different characteristics and to what degree do they differ. In this dissertation, we use concepts from network science (Cramer et al., 2015), in particular, graphs, as a visualisation tool, to have a better understanding of the patient comorbidities' prevalence and associations, and to study the interaction between medical appointments.

In the past years, network science has been used in healthcare to study several areas, including genetics, neuroscience, epidemiology and disease interaction. Networks are a simple representation of how entities connect and interact with each other, helping to reveal patterns and providing useful visualisation options. Networks may be represented as graphs, structures composed of nodes (a single point) and edges (connections between points). Edges can be directed (symmetric) or undirected (asymmetric), and can also have an associated weight, that is, a measure of the strength of a link between two nodes. Figure 2.1 shows an example of an undirected and a directed graph, where the circles represent nodes and the lines/arrows represent edges. Other important graph concepts are the node degree, which is the number of connections of a node, the path, which is a sequence of edges that leads from one node to another, and the connected component, a group of nodes within which a path exists from any one entity to any other entity. In this dissertation, we also used the clustering coefficient to better characterise the graphs, a coefficient that measures the degree to which nodes in a graph tend to cluster together (Cramer et al., 2015).

Hidalgo et al. (2009) introduces the concept of Phenotypic Disease Network (PDN), a network representation of comorbidities that can be used to study their associations, differences in phenotypes between patients and disease progression. In a PDN, nodes represent diseases and edges are weighted and represent a link between diseases. The weight can be quantified using different measures, such as co-occurrence frequency, Pearson correlation or relative risk. These networks have the ability to reveal non-obvious relationships between comorbidities that could bring important information to improve patient treatment approaches.

Figure 2.1: (a) An example of an undirected graph and (b) an example of directed graph. Adapted from Fionda and Palopoli (2011).

2.6 Survival Analysis

Survival Analysis, also called Time to Event Analysis is a branch of statistics that focus on studying the lifespan of a particular population under examination. In the case of patient phenotyping, survival analysis can be used to compare different groups, allowing to understand the risk associated with each group for a given outcome. Ahmad et al. (2014) and Gulea et al. (2021) use methods from Survival analysis to compare the obtained HF patient clusters in terms of risk association with hospitalisation and mortality outcomes.

The goal of survival analysis is to estimate the expected duration for an individual to experience an event of interest. Originally, it was developed by researchers in the medical field to study the time from treatment to death, however, survival analysis can be applied to other events besides mortality. Applications of these methods include the estimation of the lifetime of a machine, the prediction of time a customer remains with the same network operator or time until a cancer patient relapses.

As pointed out by Clark et al. (2003), standard statistical techniques are usually not applicable to survival data due to its specific characteristics. As the outcome of the study is the time between one event and another, data are rarely normally distributed, being usually skewed with the majority of events being concentrated either in early stages (e.g. time to relapse in high risk patients) or in late stages (e.g. time to death in a low risk community) of the time of the analysis. Additionally, it is common that only some individuals experience the event of interest during the time of the study. It is due to these features that survival analysis methods are necessary.

Time to event in survival analysis is always relative to a defined starting point, e.g. a disease diagnosis. All individuals must have the same starting point, which is where the time t is equal to zero and all individuals have a survival probability of one. As mentioned above, some individuals do not experience the event of interest in the time span of the study, their survival times are longer than their time in the study. The survival times for these individuals are labelled as censored. Besides the previous case, censoring also happens when an individual drops out of the study or when an individual experiences a different event that makes the follow-up non-viable.

According to Bewick et al. (2004), survival data are usually modelled and described relative to two concepts, survival and hazard. The survival function, $S(t)$, defines the probability that the event of interest has not occurred at time t, that is, that the individual survives from $t = 0$ to a time t. The hazard function, $h(t)$, is the instantaneous incident rate at time t, conditional to the subject surviving to t. It can be considered the risk of experiencing an event at time t whereas the survival function focuses on not experiencing the event. Survival relates to the cumulative non-occurrence of the event and hazard to the incident event rate.

2.6.1 Kaplan-Meier

The Kaplan-Meier method is used to estimate the survival curve, which represents the fraction of individuals who survived for a given amount of time t under the same circumstances. The survival probability is estimated from the observed survival times from both censored and uncensored individuals and relies on three assumptions:

- (1) individuals that have censored data have the same survival prospects as the uncensored;
- (2) survival probability is equal for all individuals, independently of the time they entered the study;
- (3) the event being studied happens at the specified time.

The survival probability at a specific time t is given by the number of individuals surviving divided by the number of people at risk (not including censored individuals). As events are assumed to be independent, the cumulative survival probability can be obtained by multiplying prior probabilities. Thus, the probability of an individual being alive at time t_j is computed from the probability of being alive at time t_{j-1}, the number of patients surviving until t_j, n_j and the number of events until t_j, d_j, with the equation:

$$S(t_j) = S(t_{j-1}) \cdot (1 - \frac{d_j}{n_j}) \tag{2.6}$$

where $t_0 = 0$ and $S(t = 0) = 1$. Since the value of $S(t)$ remains constant between times of events, the estimated probability is a step function that only changes value at the time of an event.

As described in Bewick et al. (2004), survival curves from different groups can be compared using the log-rank test. This statistical test is computed to test the null hypothesis that there is no difference between the survival curves from two groups.

2.6.2 Cox Proportional Hazards

Cox proportional hazards regressions analyse the effect of several variables and their relation to the survival distribution of a specified event. It is similar to a multiple regression model, allowing to differentiate survival times of particular groups of individuals while also testing for different factors. The dependent variable in this model is the hazard, which represents the probability of experiencing the event being studied. The model can be considered semi-parametric, since although it does not assume any particular survival model, the hazards ratios and the effects of predictor variables are constant over time. The hazard at a certain time t for a certain factor x, $h(t \mid x)$ is given by:

$$h(t \mid x) = b_0(t) \cdot e^{\sum_{i=1}^{n} b_i(x_i)} \tag{2.7}$$

where $b_0(t)$ is the baseline hazard and $e^{\sum_{i=1}^{n} b_i(x_i)}$ is the partial hazard. The regression coefficients b are computed by maximising the partial likelihood. The partial hazard is a time-invariant scalar factor that increases or decreases the baseline hazard depending on the value of the coefficient b_i. A positive sign for b_i indicates that the risk of an event is higher and that the prognosis for that individual is worse. A negative coefficient indicates a lower risk. The magnitude of the coefficient is also relevant to understanding the relation of the variables with the hazard. If the value of the coefficient is one, the variable does not have an impact on the hazard. However, if the value is higher than one, it will increase the hazard and if it is less than one it will decrease the hazard.

Survival analysis provides methods that allow comparing the risk for an event of interest for different groups. The Kaplan–Meier model estimates the survival curve, the log-rank test compares two groups statistically, and Cox's hazards model allows for the inclusion of additional variables.

2.7 Summary

The prevalence of multimorbidity is expected to rise in patients all around the world. Despite some studies on multimorbidity, more research is needed to understand the interaction between chronic diseases and the consequences for the patients. Only then it will be possible for healthcare systems to move from a disease-centred approach to a patient-centred one.

HF is a complex syndrome that often co-occurs with other chronic conditions. There are many possibly classifications for HF, including in terms of local of the lesion, heart functional status, and disease severity. However, there is an agreement that conventional HF classification schemes are inadequate and may fail to account for heterogeneity resulting from a wide variety of patient variables and comorbidities, which can have a significant impact on outcomes and treatment response. Thus, an improved characterisation for HF is needed.

EHRs phenotyping provides methods for the identification and characterisation of patient subgroups, which is particularly interesting in the case of complex diseases such as HF. From the different approaches analysed (rule-based, NLP, ML), clustering methods showed to be the most promising for the identification of HF subgroups. Clustering methods have been widely applied to EHRs phenotyping when the goal is to identify subgroups within a single disease. Additionally, clustering methods can incorporate mixed-type data, which is an advantage when working with heterogeneous data sources as EHRs, and there are well defined metrics to evaluate and compare clustering algorithms.

This dissertation builds on the work reviewed in this chapter, with the goal of identifying and phenotyping subgroups in HF patients with multimorbidity, using survival analysis to compare subgroups in terms of risk and introducing graphs representations as a new visualisation tool of patient subgroups.

Chapter 3

Patient Clustering Workflow

This chapter describes the patient clustering approach proposed to identify and characterise Heart Failure (HF) subgroups. Given a dataset of patients and their clinical records, the clustering model outputs the patient subgroups, their characterisation, and the association between phenotype groups and outcomes. The cluster characterisation takes into account demographics, comorbidities, laboratory values, prescriptions, and medical appointments.

Figure 3.1 illustrates the proposed workflow, to be detailed in the following sections. The first step was to select and extract HF patients with multimorbidity from the Electronic Health Record (EHR), which is described in Section 3.1. This section also describes all the features from the different domains analysed, including demographics, laboratory values, medical prescriptions, medical appointments and outcomes. The next steps, described in Section 3.2, included data preprocessing and an initial exploratory analysis, to have a better comprehension of the dataset. Section 3.3 provides an overview of the third step, the clustering analysis, including the choice of the clustering algorithm, the choice of the number of clusters k and the clustering validation indices used. Finally, Section 3.4, presents the final step, which involves characterising and comparing the clusters, using statistical tests and visualisation methods. Additionally, this section explains the application of survival analysis to study the relationship of clusters with clinical outcomes.

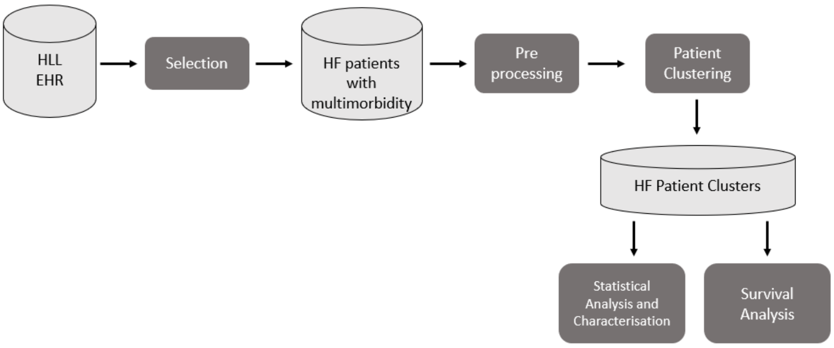


Figure 3.1: Overview of the proposed approach for the identification and characterisation of HF patient subgroups using EHR data from Hospital da Luz Lisboa (HLL)

Table 3.1: ICD-9 Codes and keywords used to identify HF patients from the HLL database. ICFEp - Insuficiência Cardíaca com Fração de Ejeção Preservada, ICFEr - Insuficiência Cardíaca com Fração de Ejeção Reduzida, IC - Insuficiência Cardíaca, ICC- Insuficiência Cardíaca Congestiva, NYHA - New York Heart Association.

ICD-9 Codes	428, 398.9.1, 402.0.1, 402.9.1, 404.0.1, 404.0.3, 404.1.1, 404.1.3, 404.9.1, 404.9.3, 425.4, 425.5, 425.6, 425.7, 425.8, 425.9
Keywords	Insuficiência cardíaca, Insuficiência cardiaca, Insuficiencia cardiaca, Insuficiencia cardíaca, ICFEp, ICFEr, IC, ICC, NYHA

Table 3.2: Feature summary of the Hospital da Luz Lisboa HF dataset.

Phenotypic Domain	**Phenotypes**
Demographics	**Age**, **Gender**
Physical Characteristics	**Body mass index (BMI)**
Lifestyle	Drug use, Alcohol use, Smoking habits
Laboratory	**Sodium**, Potassium, Bicarbonate, **Urea**, **Creatinine**, GFR, Fasting Glucose, **Hemoglobin**, **Platelet count**, **Red Cell Distribution Width (RDW)** , **N-terminal-pro-BNP (NT-proBNP)**, Ferritin, Uric Acid, Sedimentation Rate
Comorbidities	**Ischemic Cardiomyopathy**, **Hypertension**, **Diabetes**, **Atrial Fibrillation**, **Cerebrovascular Disease**, **Valvular Disease**, **Chronic Kidney Disease**, **Anaemia**, **Chronic Obstructive Pulmonar Disease**, **Obesity**
Patient complexity	**Number of non-chronic diseases**, **Number of chronic diseases**, **Number of ICD-9 codes**, **Number of consultations**

3.1 Data

The dataset used to develop the pipeline was generated from an initial population of 54 827 patients, with an observation period between January 2007 and August 2021. From this initial pool, 3745 HF patients with multimorbidity were identified using the International Classification of Diseases (ICD)-9 for HF and heart disease, or by identifying associated keywords in their medical records (see Table 3.1).

For the clustering analysis, relevant features were selected based on a literature review and guidance from the HF specialist at HLL. These features consisted of clinical variables, demographics, physical characteristics, laboratory data and the most common comorbidities associated with HF, amounting to a total of 35 features. The specific features for each phenotypic domain can be seen in Table 3.2. Lifestyle-related features and gender were provided as text fields, comorbidities as binary variables based on the presence or absence of the disease, and all others as numeric. Besides the features used for clustering, we also extracted the date of HF diagnosis and gathered data on other comorbidities, prescriptions (medications ordered for each patient), medical appointments, and clinical outcomes, namely hospitalisations, emergency admissions and mortality. Data for prescriptions were extracted from a different information system than the rest of the clinical data, and so these data are relative to only a part of the total time period, from January 2012 until August 2021 (9 years and 8 months). Data for medical appointments and outcomes are relative to the full period under analysis, from January 2007 until August 2021 (14 years and 8 months). For outcomes hospitalisation and emergency admissions, it was possible to obtain the entries of all instances that occurred and the respective date by patients. For the outcome mortality, it was not possible to obtain information regarding the date of death.

3.2 Preprocessing

Prior to the clustering, it was necessary to preprocess the dataset obtained from the EHR extraction. Categorical features (Gender) were converted into numeric binary features and features with a prevalence lower than 2% in the cohort were removed (Drug use). Features with a percentage of missing values higher than 40% were deleted: Alcohol use, Smoking habits, Glomerular Filtration Rate (GRF), Fasting glucose, Potassium, Ferritin and Bicarbonate. The distribution of missing values in the dataset can be seen in Figure 3.2. Of the remaining features, the ones containing missing values were the following: Body Mass Index (BMI) (38.53%), Sodium (15.22%), Urea(15.54%), Creatinin (15.08%), Hemoglobin (13.69%), RDW (13.69%), Platelet count (13.69%), NT-proBNP (28.17%). According to Waljee et al. (2013), the two methods that resulted in the least imputation error and prediction difference when applied to a dataset containing laboratory data were missForest and multivariate imputation by chained equations (MICE). Missing values were imputed using Python's function *Iterative Imputer*, which is based on the Multivariate Imputation by Chained Equations (MICE) method. The MICE method, as described in Azur et al. (2011), models the missing values of each feature as a function of other features. To do that, at each step, one of the feature columns is designated output y and the rest of the feature columns are designated as inputs x. To cover for possible coding errors, the feature Anemia was defined based on the value of Hemoglobin (Hemoglobin $<$ 12 for women and Hemoglobin $<$ 13.5 for men, according to Yilmaz and Shaikh (2020)) and the feature obesity feature was defined based on the value of the feature BMI (BMI $>$ 30 according to OECD/WHO (2020)). Continuous features were normalised to have a mean of 0 and a standard deviation of 1 (using Python's function *StandardScaler*). Categorical binary features were scaled from $\{0, 1\} \rightarrow \{-0.5, 0.5\}$. After preprocessing the total number of features used for clustering was 25. The features are identified in Table 3.2 in bold.

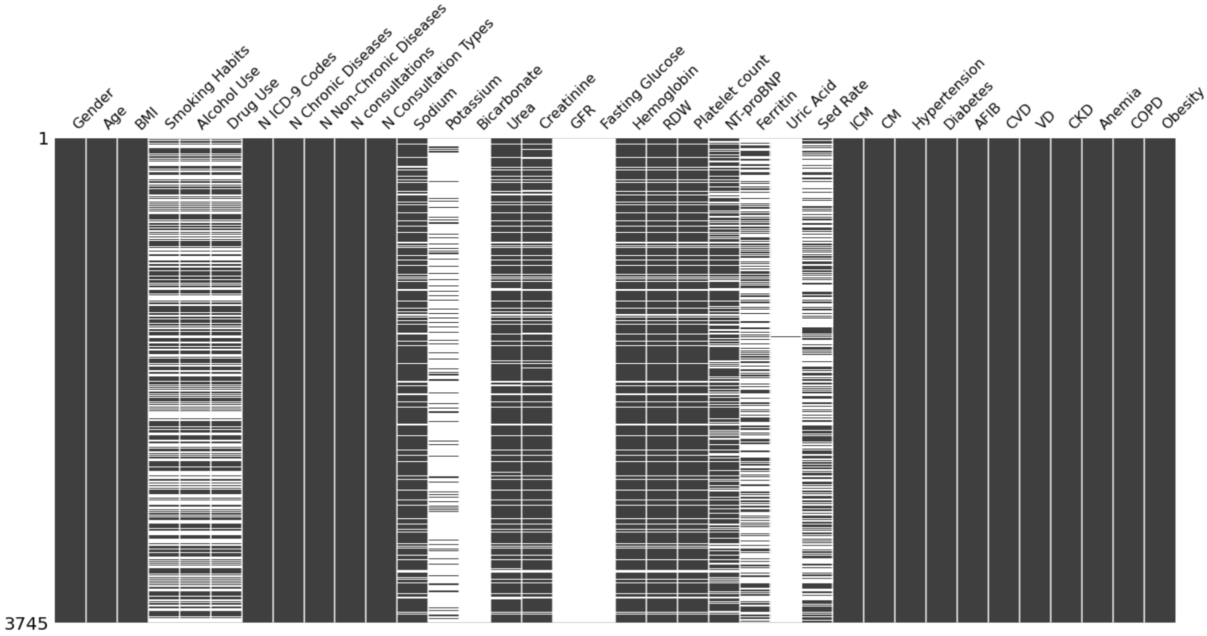


Figure 3.2: Missing values in initial HF dataset from HLL

In addition to the preprocessing, an Exploratory Data Analysis (EDA) was also performed. EDA uses summary statistics and graphical representations to analyse datasets and get a better comprehension of the data. Histograms and boxplots were used to visualise the distribution of continuous variables and barplots were used to visualise the distribution of categorical variables, such as the distribution of comorbidities or medical appointments. Continuous data were described using median (25th -75th quantile) and categorical data using percentages. A comparison stratified by gender was also performed in order to understand if there were significant differences in female and male patients.

3.3 Patient Clustering

To apply clustering to the HF dataset from HLL, it was first necessary to determine which clustering algorithm and possible complementary techniques to use, and how to evaluate the clustering. It also was necessary to take into account that the dataset was composed of both numerical and categorical data. Several clustering algorithms were tested to understand which one would be more suitable for the HF data. Based on the literature regarding clustering mixed-type data, presented in Section 2.5.1, and on HF clustering, presented in section 2.5, the following combinations of methods were tested for this dataset:

1. Gower's distance matrix together with Ward's Agglomerative Hierarchical Clustering

2. Dimensionality reduction followed by Ward's Agglomerative Hierarchical Clustering

3. Dimensionality reduction followed by K-Means

The dimensionality reduction method chosen was Factor Analysis of Mixed Data (FAMD), a principal component method specific to analyse quantitative and qualitative variables (Husson et al., 2008).

3.3.1 Clustering Evaluation

To evaluate the resulting clusters, three different validity indices were used. Considering that clustering HF patients is an unsupervised analysis, there are no ground truth labels. Hence, only internal validity indices, which depend uniquely on the data being clustered and on the resulting labels, could be used. The used indices were Silhouette Score, Calinski-Harabasz, and Davies-Bouldin (see Section 2.4.1).

After computing the three indices for the different clustering algorithms and for different values of k, the method to use was chosen using a majority vote, meaning that the algorithm and k that performed best in at least two of the indices was chosen. Additionally, as in Ahmad et al. (2014), a minimum of $N > 375$ was also defined to promote stability and ensure that there was not any cluster with less than 10% of the total population.

3.4 Statistical Analysis and Characterisation of Obtained Clusters

After choosing the number of clusters and obtaining the phenotype groups, it was necessary to characterise the patient subgroups, evaluate if they had statistically significant differences, and determine adequate methods to help visualise the clusters. Following the work by Gulea et al. (2021), demographic, clinical and laboratory characteristics were compared between groups using Chi-squared tests for categorical variables and Kruskal-Wallis test for continuous variables, computing the respective p-values. Clusters were characterised according to age, gender, and most predominant comorbidities.

To have a better visualisation of the clusters comorbidities prevalence, and associations, we computed a graph representation of each cluster. In this case, each graph's node represents a disease in the cluster and an edge represents a co-occurrence of the two nodes (diseases) connected by that edge. The graphs were created with Python's *NetworkX*[1] package and *Gephi*[2] for visualisation. The settings were adjusted so that node size was proportional to the number of connections to other nodes (node degree) and edge thickness was proportional to disease co-occurrence prevalence (edge weight). Co-occurrences (edges) with a co-occurrence prevalence lower than 2% were discarded to declutter the visualisation. The graph representation of the subgroup comorbidities provides a better understanding of which diseases co-occur more frequently and which diseases have the most connections with other diseases.

The obtained patient subgroups were also discussed with the HF specialist to understand if they were clinically meaningful, and comparable to previous HF clustering studies.

3.4.1 Prescriptions

We collected information on medication prescriptions, which included prescription dates, the medical speciality in which the prescription was appointed, and the commercial and common names of the medication prescribed. Following treatment guidelines from the European Society of Cardiology in European Society of Cardiology (2016) for HF and advisory from the HF specialist, we chose the most relevant medication groups to be analysed. These groups included HF specific medications, such as Beta-blockers, Angiotensin-Converting-Enzyme Inhibitors (ACEi) and Angiotensin Receptor Blockers (ARB), but also medications for the comorbidities found in the dataset, such as Metformin for type-2 diabetes or Bronchodialaters for Chronic Obstructive Pulmonary Disease (COPD) (see description of all medication groups considered, the medications included for each one and a brief overview of their use in Table 3.3). We compared the prevalence of the medication groups in each cluster was compared, having into account the most common comorbidities in the clusters. The distribution of medication groups between clusters was compared with Chi-squared tests, computing the respective p-value.

[1] https://networkx.org/

[2] https://gephi.org/

Table 3.3: Medication groups considered and medications included in each one. ACEi - angiotensin–converting enzyme inhibitor; MRA - Aldosterone receptor antagonists; DPP4i - Dipeptidyl peptidase-4 inhibitor; ARB - angiotensin receptor blockers; ARNi - Angiotensin Receptor-Neprilysin Inhibitors; SGLT2i - Sodium-glucose cotransporter 2 inhibitors; GLP-1-Glucagon-like peptide-1

Medications groups	**Common usage**	**Medications included**
Beta-Blockers	Reduce blood pressure, often used for HF, AF, Angina	Atenolol, Bisoprolol, Carvedilol, Metoprolol, Nebivolol, Propanolol
ACEi\ARBs	Reduce blood pressure, often used for HF, Coronary Heart Disease	Lisinopril, Captopril, Enalapril, Perindopril, Ramipril, Trandolapril, Candesartan, Irbesartan, Losartan, Olmesartan, Telmisartan, Valsartan
ARNi	Reduce blood pressure, often used for HF, Chronic Kidney Disease	Sacubitril/Valsartan
Diuretics	Lower blood pressure	Clorotalidona, Hidroclorotiazida, Indapamida, Furosemide, Metolazona, Torasemido
MRA	Manage treatment-resistant forms of hypertension, often used for HFpEF	Eplerenona, Espironolactona
Digoxin	Treat abnormal heart rhythms, often used for HF, AF	Digoxin
Anticoagulants	Prevent blood clots	Acenocumarol, Apixabano, Dabigatrano, Edoxabano, Rivaroxabano, Varfarina
Antiplatelets	Prevent blood clots	Acetylsalicylic acid, Clopidogrel, Ticagrelor
Statins	Lower the level of low-density lipoprotein (LDL) cholesterol	Atorvastatin, Fluvastatina, Pitavastatina, Pravastatina, Sinvastatina, Rosuvastatina
SGLT2i	Lower blood glucose level, often used for type 2 diabetes	Dapagliflozina, Empagliflozina
Ivabradine	Lower heart rate, often used for HF	Ivabradine
Metformin	Lower blood glucose level, often used for type 2 diabetes	Metformin
Insulin	Maintain normal blood glucose levels, used for type 1 diabetes	Insulin
GLP-1 antagonists	Lower blood glucose level, often used for type 2 diabetes	Liraglutido, Exenatida, Dulaglutida, Semaglutido
Sulfonylureas	Lower blood glucose level, often used for type 2 diabetes	Glimeperida, Gliclazida, Glibenclamida
DPP4i	Lower blood glucose level, often used for type 2 diabetes	Linagliptina, Sitagliptina, Vildagliptina
Levothyroxine	Treat an underactive thyroid gland (hypothyroidism)	Levothyroxine
Antiarrhythmics	Treat abnormal heart rhythms	Amiodarona, Propafenona
Inhalers Bronchodilator	Dilate the lungs' airways, often used for Asthma and COPD	Salbutamol, Salmeterol, Indacaterol, Formoterol
Inhaled Corticosteroids	Reduce inflammation in the lungs, often used for Asthma and COPD	Beclometasona, Fluticasone
Inhaled Anticholinergics	Block the action of acetylcholine, often used for COPD, bladder disorders	Ipratromium bromide, Tiotropium bromid, Aclidinio
Hematinic factors	Treat or prevent low blood levels of iron	Folic Acid , Ferrous Sulfate, Cianocobalamina

3.4.2 Medical appointments

Data on medical appointments were also gathered to better characterise the clusters and investigate which medical specialities were more predominant in each cluster. The information on medical appointments included the date and medical speciality of the appointment, and so it was possible to analyse not only the prevalence of each medical speciality, but also the most common sequences of appointments. A graph representation of medical appointments was then computed using the temporal sequence of each patients' medical appointment history. In this representation, a directed graph was used, in which the node size was proportional to the total number of connections to other nodes (previous or following appointment), and the edge thickness was proportional to the occurrence of that sequence of medical appointments. Only medical appointments with a prevalence in the dataset higher than 2% were considered for this analysis.

3.4.3 Outcomes

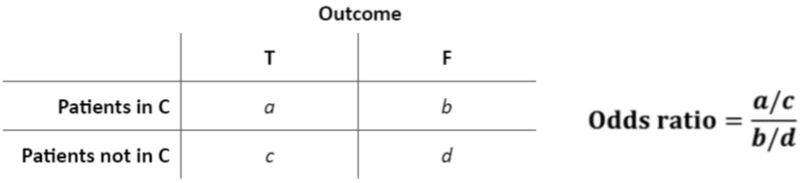
Finally, we examined hospitalisations, emergency admissions to the hospital and mortality. A hospitalisation was considered to be any admission to the hospital requiring an overnight stay. Hospitalisations are often planned, but can also occur after appointments, examinations or emergency admissions. An emergency admission, in this case, represents an unplanned, often urgent admission, which occurs when a patient is admitted at the earliest possible time. For the outcomes of hospitalisation and emergency admission, we computed the incidence per year, the percentage of patients that had an incidence in the first year after HF diagnosis, and the percentage of patients that had an incidence at least one during the time period analysed. For the outcome mortality, we computed the incidence in each cluster. The values obtained were compared using Chi-squared tests and computing the respective p-values.

Additionally, we computed the Odds Ratio (OR) for all outcomes, for each cluster in relation to all other clusters, comparing the patients belonging to the cluster with all patients not in that cluster (see Figure 3.3). OR, as defined in Szumilas (2010), measure the association between an exposure and an outcome, comparing the chance of an event occurring in the absence of a certain exposure against the chance of that outcome occurring in the presence of that exposure. A value of OR equal to 1 means that the exposure does not affect the odds of the outcome, a value higher than 1 indicates that the exposure is associated with higher odds of the outcome and a value lower than 1 indicates lower odds. In this case, the exposure considered is belonging to a certain cluster. The formula used to compute the OR can be found in Figure 3.3.

It is important to note that the association computed with OR do not translate into causality, and so the value of OR provide only a measure of association.

3.4.4 Survival Analysis

We conducted a Survival Analysis to analyse the relationship between the clusters and the evolution of the outcomes since diagnosis and to understand if belonging to a certain disease subgroup would mean the patient is at a higher risk of a certain outcome in the future. While OR are useful to compute a



	Outcome	
	T	F
Patients in C	*a*	*b*
Patients not in C	*c*	*d*

$$\textbf{Odds ratio} = \frac{a/c}{b/d}$$

Figure 3.3: Calculation of odds ratio between patients in one cluster C and patients not in C for an outcome, given the number of patients for each outcome in each group.

measure of association between the clusters and the outcomes, Survival Analysis provides a direct comparison between clusters, through the survival curves and Hazard Ratio (HR). As temporal information regarding the outcomes was only available for outcomes of hospitalisation and emergency admission, the outcome mortality was not included in this analysis. Two separate analyses were performed for the outcomes of hospitalisation and emergency.

To perform Survival Analysis, it is necessary to define the starting point that will be common to all patients. In this study, the starting point, or t_0, was defined as the moment of diagnosis, and all intervals for the survival analysis were calculated in relation to that moment. The first step was to compute the time intervals between diagnosis and the occurrence of an outcome for all patients. Patients that did not experience the outcome were labelled as censored. Following this and using Python's package *Lifelines* [3], Kaplan-Meier curves and Cox proportional regression models were computed. Kaplan-Meier curves were computed for each outcome individually and were stratified per cluster, with differences between groups tested using the log-rank test. For Cox proportional regression, following what was done by Shah et al. (2014), three different models were used for each outcome: an unadjusted model that only took into account the clusters, a second model adjusted for age and gender and a third model adjusted for age, gender and the laboratory value NT-proBNP, which is often used as a risk marker for HF. HR from Cox regression models are presented in relation to the lowest risk cluster (determined by the lowest percentage of outcomes).

Kaplan-Meier allows visualising the progress of the survival curves of the different clusters, providing a comparison in survival time between clusters. Cox proportional regression allows to compute HR, which measures how belonging to a certain cluster changes the rate of experiencing an outcome. Both of these methods are useful to compare subgroups of patients in terms of the risk they are exposed to and disease severity.

3.5 Summary

This chapter detailed the architecture of the proposed workflow to identify and characterise HF patient subgroups using EHR data.

For this purpose, we selected and preprocessed data from HF patients with multimorbidity from the EHR database of HLL. The features used for clustering included demographics, comorbidities and lab-

[3] https://lifelines.readthedocs.io/en/latest/

oratory values, and the most appropriate clustering algorithm for this dataset was determined using clustering evaluation metrics. To characterise the resulting clusters, statistical analyses were performed to the features used for clustering and data from medical prescriptions, medical appointments and outcomes were also analysed. Graph representations of the clusters comorbidities and their associations, and of medical appointments transitions were used to visualise the clusters and compare their complexity. To determine the association of the clusters with the outcomes analysed, a survival analysed was performed.

With the proposed workflow, the objective is to obtain the most detailed phenotype possible for the HF patient subgroups in the different domains analysed, and to understand to what degree are the phenotypes related to disease severity and risk of hospitalisation and emergency admissions.

Chapter 4

Results and Discussion

This chapter describes the evaluation of the Heart Failure (HF) patients subgroups obtained by applying the proposed patient clustering method to data from Hospital da Luz Lisboa (HLL). Section 4.1 presents a statistical characterisation of the HF dataset. Section 4.2 discusses the process for selecting the optimal clustering algorithm and number of clusters k. Section 4.3 provides a characterisation of the obtained clusters, describing all features analysed and giving a clinical perspective on the results, including a survival analysis performed to understand the association between the clusters and outcomes.

4.1 Characterisation of the Heart Failure dataset

HF patients are usually complex patients with advanced age and with a high number of other diseases, which is visible in the dataset used in this dissertation. The HLL dataset includes 3745 records of HF patients with multimorbidity. The median age is 82 years (inter-quartile range 73-88), the number of female patients is 1979 (52.84%), the number of male patients 1766 (47.16%), and the median Body Mass Index (BMI) is 24.8, as can be observed in Figure 4.1. The median number of chronic diseases is 5 (inter-quartile range 3-7) with approximately 40% of the population having between 3 and 5 comorbidities and approximately 30% having 6 to 8 comorbidities.

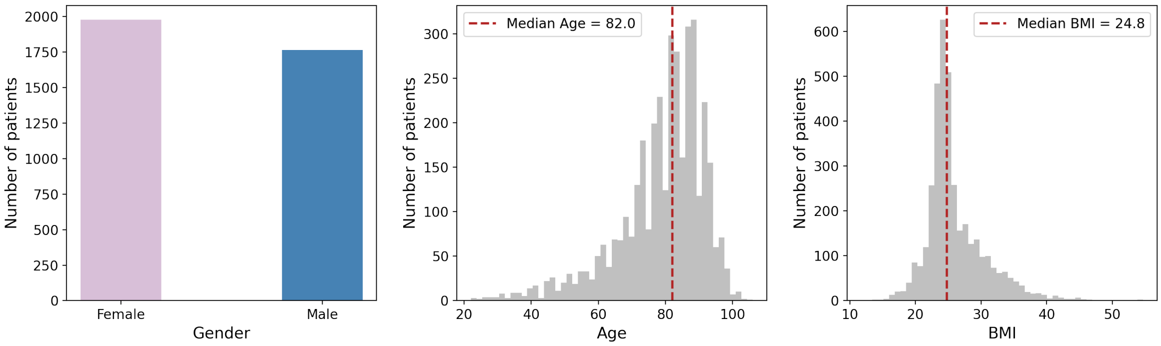


Figure 4.1: Gender, Age and BMI distribution of the dataset of HF patients with multimorbidity.

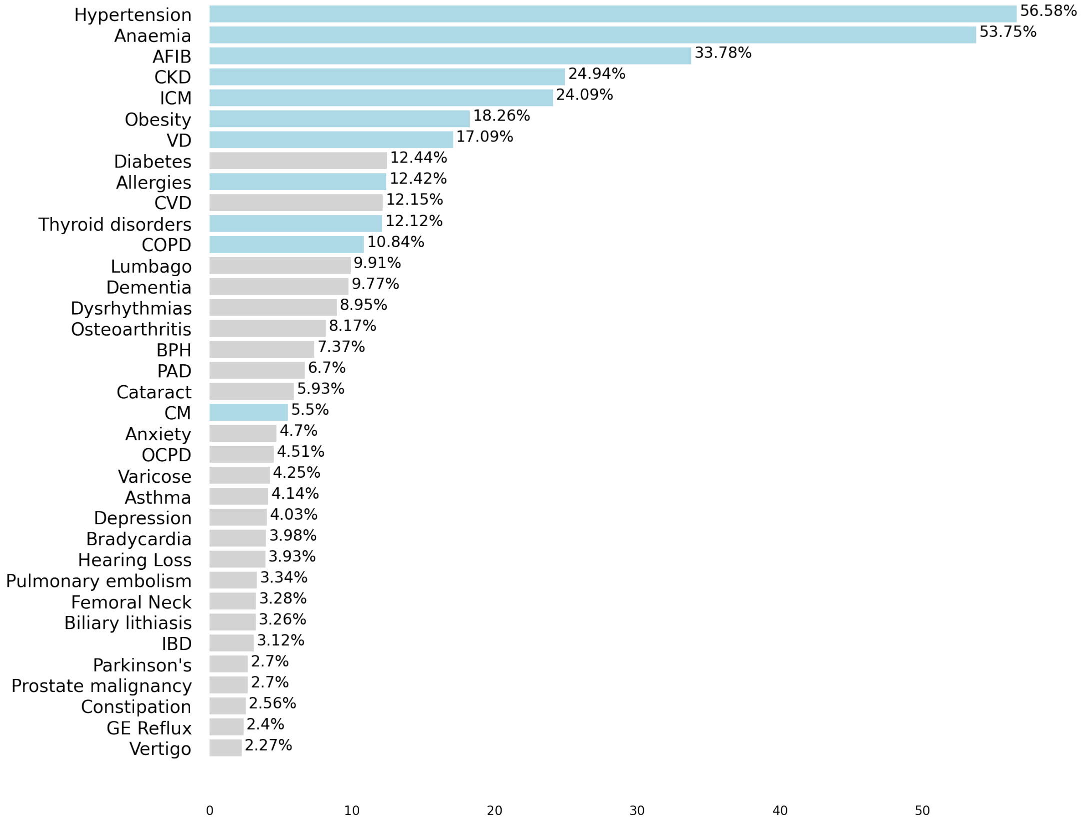
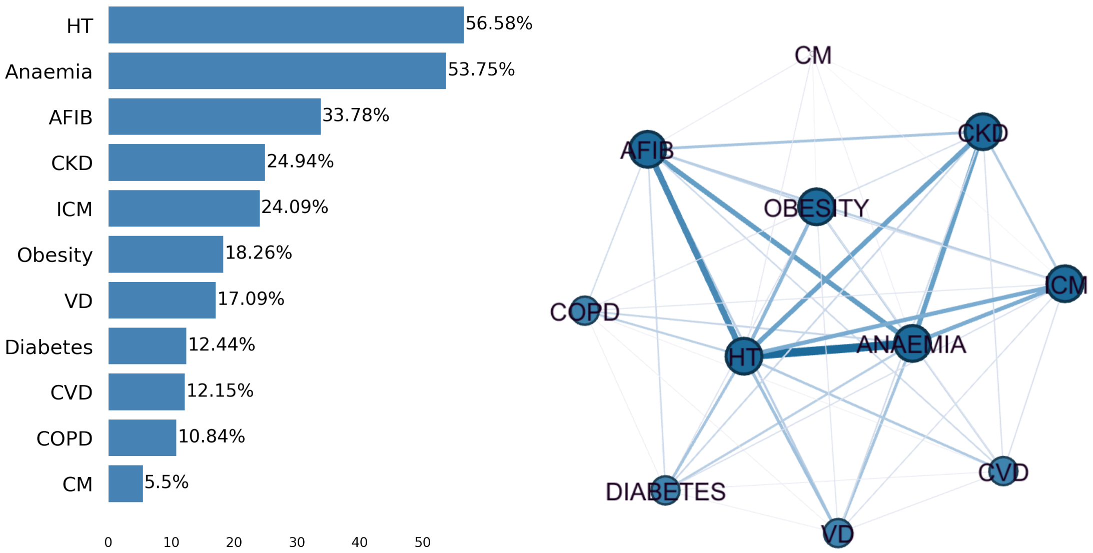


Figure 4.2: Prevalence of all comorbidities in the HF dataset.

Besides the diseases used for clustering, the prevalence of other diseases in the dataset was also analysed in order to have a more complete profile of the patients under study. Figure 4.2 presents the frequency of all diseases found in the dataset, with a prevalence higher than 2%. Diseases used for clustering are represented in blue. The comorbidities used for clustering (in blue) were chosen under the advisory of the HF specialist. These diseases correspond to the ones that have a higher co-occurrence with HF and are also the seven most prevalent diseases in the dataset from HLL. The most common comorbidity in the dataset is Hypertension (HT) (56.58%), followed by Anaemia (53.75%) and Atrial Fibrillation (AFIB) (33.78%). Other diseases with a high prevalence are Chronic Kidney Disease (CKD) (24.94%) and Ischaemic Cardiomyopathy (ICM) (24.09%). Observing other comorbidities with a high prevalence in the dataset, there are several diseases in the dataset that are expected to be found in an older population such as the one being studied, as for example Dementia, Osteoarthritis, Cataract and Hearing Loss. Other diseases found in the dataset are cardiac diseases that can also be related to HF, such as Dysrhythmias, Peripheral Arterial Disease (PAD) and Bradycardia. Some mental disorders can also be found with a lower prevalence, namely, Anxiety, Depression and Obsessive Compulsive Personality Disorder (OCPD).

In addition to assessing the prevalence of each comorbidity, a graph representation of the comorbidities was also computed (see Figure 4.3) in order to better understand the interaction and the complexity of these comorbidities. In the graph, nodes represent diseases and edges represent the co-occurrence of diseases. The size of the nodes is proportional to the number of other nodes it is connected to (node



(a) Prevalence of comorbidities.

(b) Graph representation of comorbidities.
Nodes = 11, edges = 51, average degree = 9.3,
average clustering coefficient = 0.95

Figure 4.3: Prevalence and graph representation of comorbidities used for clustering in the HF dataset. In the graph a node represents a disease and its size is proportional to the node degree. An edge represents a co-occurrence of two diseases and its width is proportional to the prevalence of the co-occurrence in the dataset. CM-Ischaemic Cardiomyopathy, HT-Hypertension, AFIB-AtrialFibrillation, CVD-Cerebrovascular Disease, VD-Valvular Disease, CKD-Chronic Kidney Disease, COPD-Chronic Obstructive Pulmonary Disease.

degree), and the width of the edge is proportional to the prevalence of the co-occurrences. Using this type of visualisation we are able to observe which diseases co-occur more frequently while also suggesting possible triadic co-occurrences. In this population, HT, AFIB and ICM are the diseases with higher prevalence and the ones that co-occur more frequently with other diseases. The thickness of the edges makes it possible to verify that HT and Anaemia, and HT and AFIB occur frequently together. CKD and ICM also show a high co-occurrence with HT and Anaemia. We will see in the clustering step that these diseases and their co-occurrences play a determinant role when obtaining clusters of HF patients.

In Figure 4.4, we also present the distribution of laboratory parameters used for the clustering analysis. Creatinine and Urea are two indicators of kidney function. High values of these markers are warning signs for possible malfunction or failure of the kidneys, which can be related to the prevalence of CKD in the dataset. The high percentage of Anaemia in the dataset is also reflected in the Hemoglobin histogram, where a high percentage of patients is below the reference value. Red Cell Distribution Width (RDW) is a measure of how equal red blood cells are in size. A high value indicates that red blood cells are being produced in different sizes which may indicate an issue with red blood cell production or survival. B-type Natriuretic Peptide (BNP) is a hormone produced by the heart and N-terminal-pro-BNP (NT-proBNP) is a prohormone that is released from the same molecule that produces BNP. Both BNP and NT-proBNP are released in response to changes in the heart. High levels of NT-proBNP are usually associated with the development of HF or with a worsening of the condition.

Another domain analysed to have a better characterisation of the population was medical appoint-

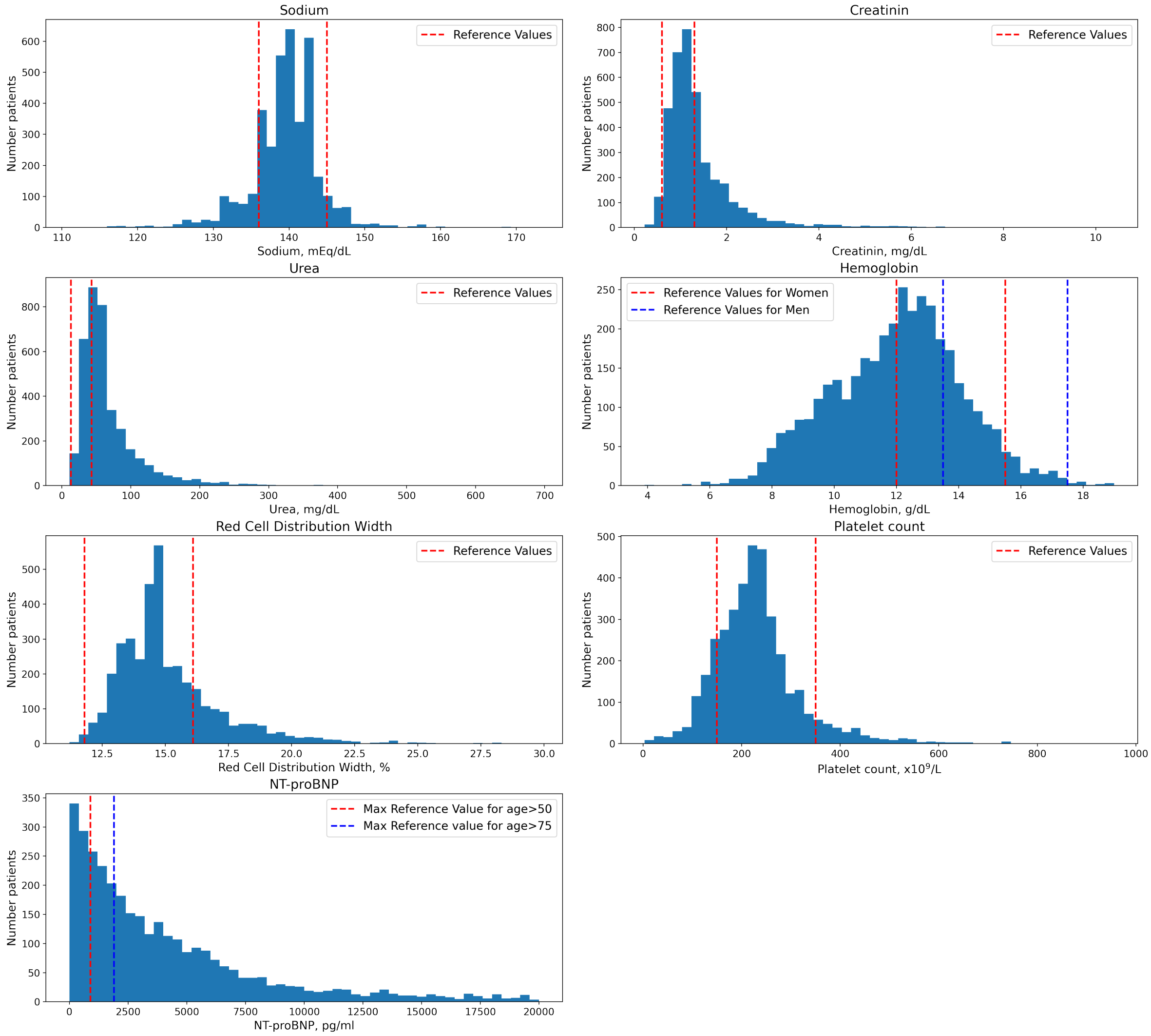
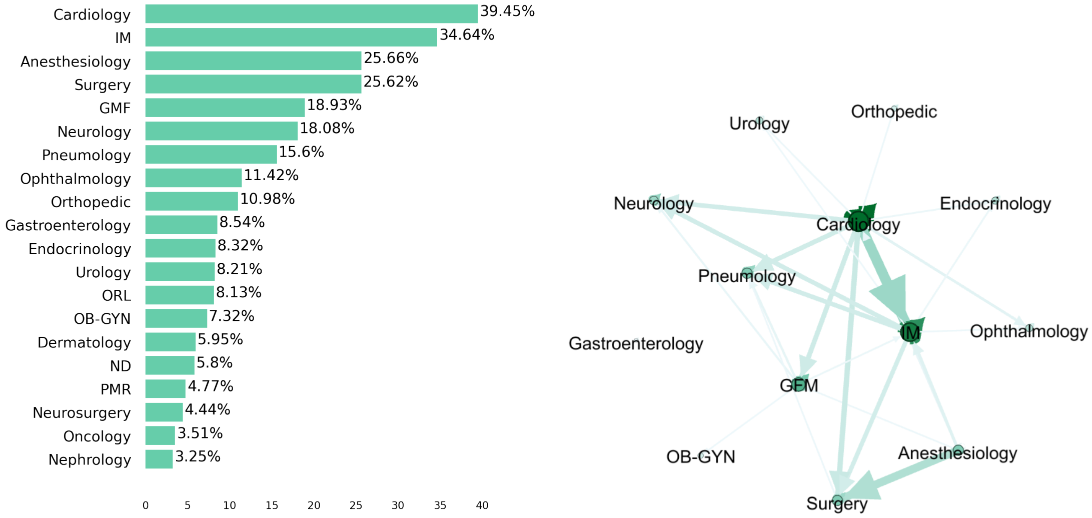


Figure 4.4: Histogram for laboratory variables for the HF dataset. Reference values are marked in red and blue dashed lines.

ments. Data from medical appointments were available for 2705 patients during a time period of 14 years (approximately), corresponding to 72.23% of the dataset. Figure 4.5a presents the distribution of the speciality of medical appointments registered for HF patients, or, in other words, the percentage of patients that had a medical appointment from that medical speciality during the 14 years analysed. The most common medical specialities in medical appointments of the HF patients were Cardiology (39.45%), Internal Medicine (34.64%), Anesthesiology (25.66%) and Surgery (25.62%), the last two being highly interconnected since a Surgery appointment is frequently preceded by an Anesthesiology appointment. The percentage of patients that had a Cardiology appointment is quite low for what could be expected in a dataset of HF patients, considering that patients with HF benefit from regular follow-up and monitoring (European Society of Cardiology, 2016). There are several reasons why this might happen, one being that follow-up appointments are provided by other specialities, such as Internal Medicine or General and Family Medicine, or that patients are also followed outside HLL. Other common medical specialities include Neurology and Pneumology. The first may be linked to the high median age of the



(a) Prevalence of medical appointments.

(b) Graph representation of medical appointments. Nodes = 12, edges = 37, average degree = 2.9, average clustering coefficient = 0.36

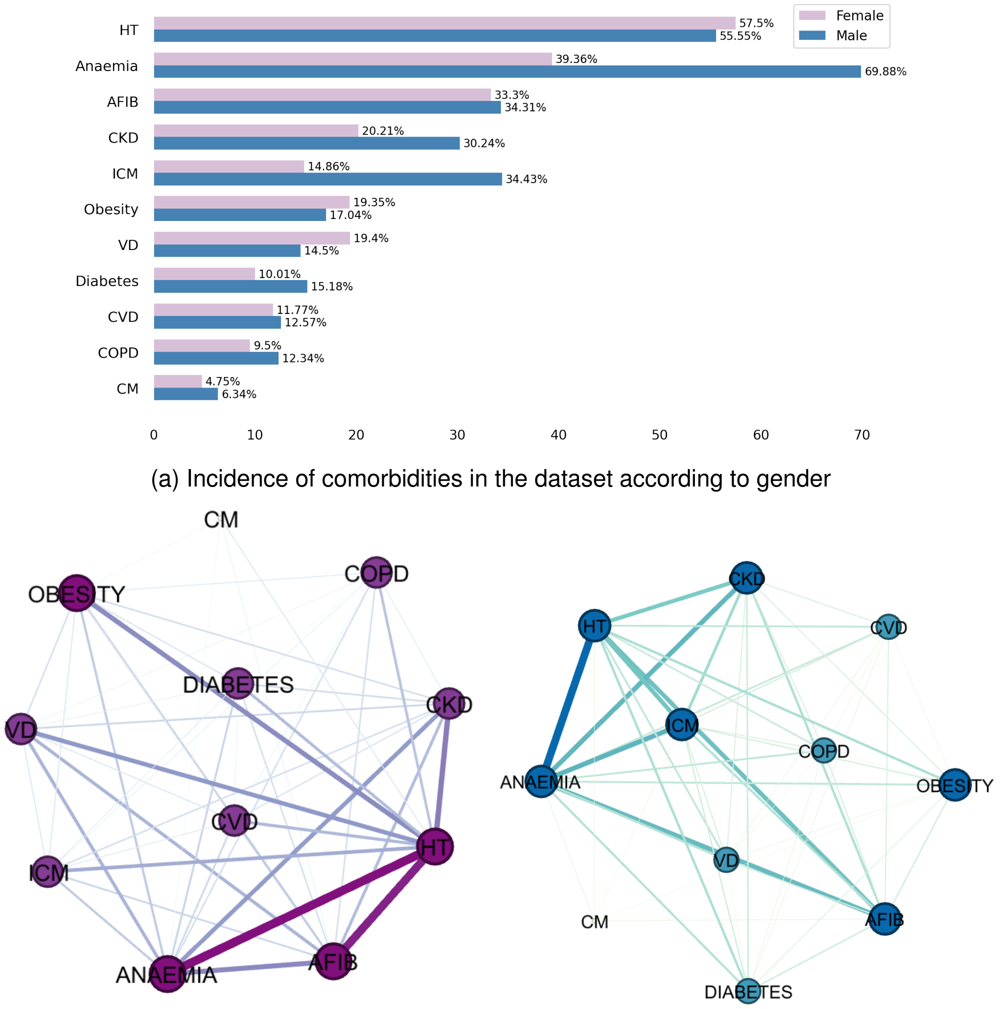
Figure 4.5: Prevalence and graph representation of medical appointments by speciality in the HF dataset. In the graph a node represents a medical appointment speciality and its size is proportional to the node degree. An edge represents a transition from one medical appointment to another and and its width is proportional to the prevalence of that transition in the dataset. ND- Nutrition and Dietetic, PMR – Physical Medicine and Rehabilitation, GFM – General and Family Medicine, IM – Internal Medicine.

population, and the second to the influence of HF in the pulmonary system, also represented by the prevalence of Chronic Obstructive Pulmonary Disease (COPD) in the dataset. Figure 4.5b shows the graph representation of the medical appointments. The graph representation used for medical appointments is a directed graph where nodes represent medical specialities and edges represent transitions of medical appointments. The size of the node is proportional to the number of other nodes it is connected to (either inward or outward connections) and the size of the edge is proportional to the number of patients that underwent the corresponding transition at least once. It is possible to observe that the bigger nodes correspond to the medical specialities with the highest percentage of appointments, Cardiology and Internal Medicine. The most common transitions between medical specialities include Cardiology to Internal Medicine and Anesthesiology to Surgery. The medical specialities Cardiology also has connections to many other medical specialities, indicating that following a Cardiology appointment patients are often directed to other specialities, possibly to treat other comorbidities.

4.1.1 Comparison of Female and Male Patients

A comparison between female and male patients was also performed to understand if there were relevant differences in the baseline characteristics considered. The median age for female patients is 80 (inter-quartile range 71-87) and for male patients 83 (inter-quartile range 75-89), and the median BMI is 25.1 for female patients and 24.4 for male patients.

Figure 4.6 shows the prevalence of the comorbidities considered for clustering and their graph representation, in female and male patients. Female patients present a higher prevalence of diseases HT, Obesity and Valvular Disease (VD). The prevalence of all other diseases is higher in male patients. For the diseases where female patients have a higher prevalence the difference is below 5%, whereas for Anaemia, CKD and ICM, where male patients have a higher prevalence, the difference is above 10%. The graph representation also allows noting differences in these patients. Female patients have a high co-occurrence of AFIB and HT, and Anaemia and HT. Male patients also have a high co-occurrence of Anaemia and HT, but also of other connections such as Anaemia and ICM, Anaemia and CKD and HT and ICM, also have a high co-occurrence. It is also interesting to note how a disease with a similar prevalence in both groups, such as AFIB, is differently associated with other diseases in each of the groups. Additionally, it is possible to observe that diseases like Obesity and VD play a more central role in the female graph representation, whereas diseases like Diabetes and COPD are more highlighted in the male graph.



(a) Incidence of comorbidities in the dataset according to gender

(b) Female graph.
Nodes = 11, edges = 49, average degree = 8.9, average clustering coefficient = 0.95

(c) Male graph.
Nodes = 11, edges = 51, average degree = 9.3, average clustering coefficient = 0.95

Figure 4.6: Prevalence and graph representation of comorbidities used for clustering by gender. A node represents a disease and its size is proportional to the node degree. An edge represents a co-occurrence of two diseases and its width is proportional to the prevalence of that co-occurrence. CM-Ischaemic Cardiomyopathy, HT-Hypertension, AFIB–Atrial Fibrillation, CVD–Cerebrovascular Disease, VD-Valvular Disease, CKD–Chronic Kidney Disease, COPD–Chronic Obstructive Pulmonary Disease

Regarding medical appointments, it is also possible to observe some differences between female and male patients. Figure 4.7 shows the distribution of attendance of medical appointments by medical speciality during the time period analysed (14 years). Male patients have a higher percentage of Cardiology appointments and also a higher percentage of Surgery and Anesthesiology appointments, whereas female patients have a higher percentage in the majority of other medical specialities. Female patients also have a higher interaction between medical specialities, that is, a higher number of medical appointments transitions, as it is possible to see in the female graph in Figure 4.8. Considering the graph characteristics, the female graph has a higher number of edges (37 vs. 29) and also a higher average degree (3.1 vs. 2.4) than the male graph. The average degree indicates the average number of connections of each node, in this case, the average number of different medical appointment transitions.

In terms of outcomes, Figure 4.9 shows the results for the distribution of the number of hospitalisations and the number of emergency admissions per year. For both hospitalisations and emergency admissions, the median is below 1, meaning that 50% of the patients experienced less than one outcome per year, during the time period analysed. For the outcome hospitalisation, the distribution is very similar for female and male patients. For the outcome of emergency admission, it is possible to observe that female patients have a slighter higher median (0.3 vs 0.2) but the distribution is also very similar.

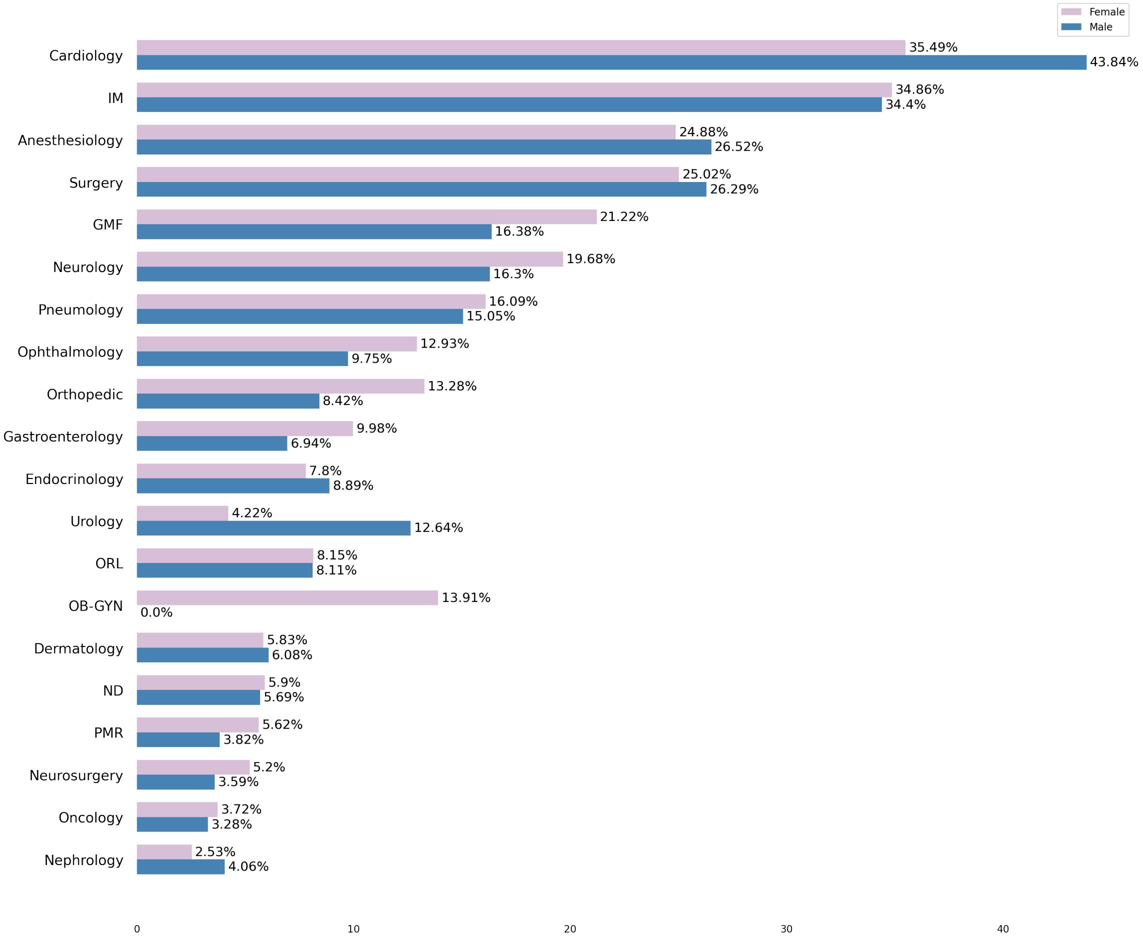
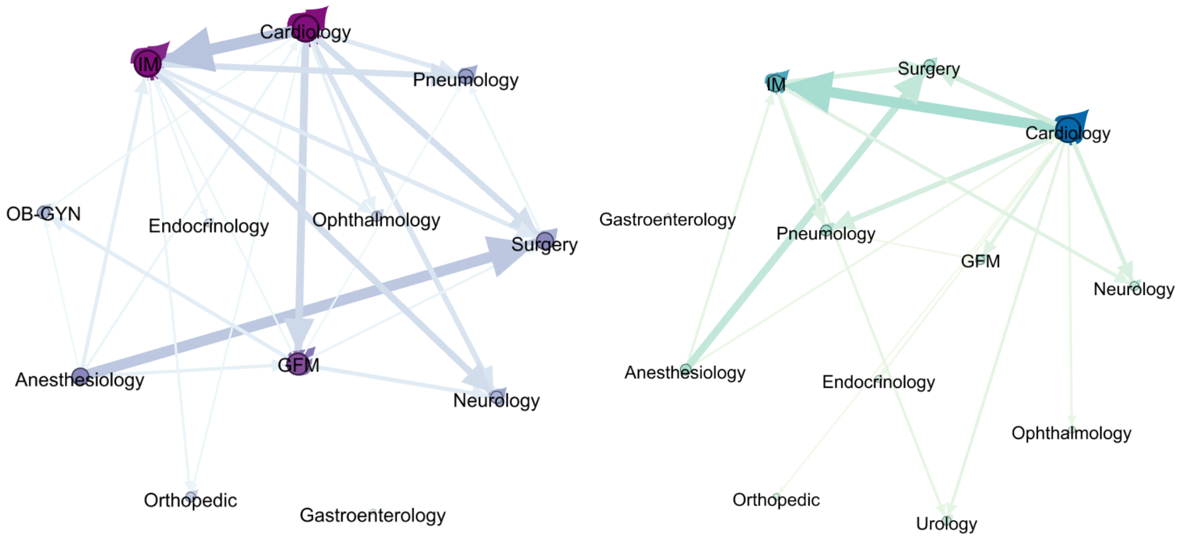


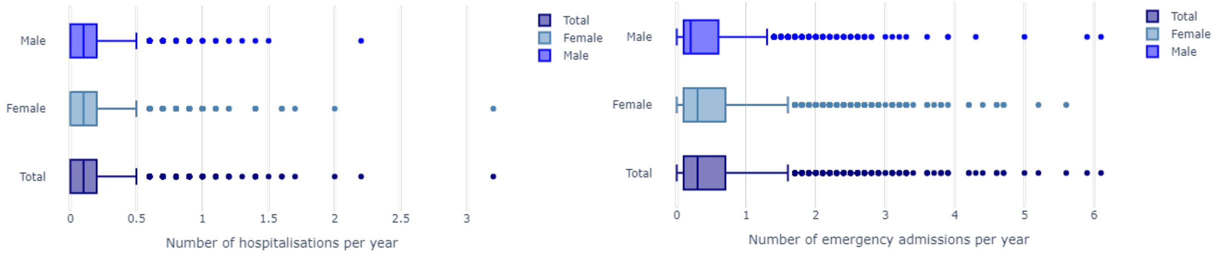
Figure 4.7: Medical appointments by gender. ND- Nutrition and Dietetic, PMR - Physical Medicine and Rehabilitation, GFM - General and Family Medicine, IM - Internal Medicine.



(a) Female graph.
Nodes = 12, edges = 37, average degree = 3.1,
average clustering coefficient = 0.46

(b) Male graph.
Nodes = 12, edges = 29, average degree = 2.4,
average clustering coefficient = 0.30

Figure 4.8: Graph representation of medical appointments by gender in the HF dataset. A node represents a medical appointment speciality and its size is proportional to the node degree. An edge represents a transition from one medical appointment to another and and its width is proportional to the prevalence of that transition in the dataset. GFM - General and Family Medicine, IM - Internal Medicine



(a) Number of hospitalisations per year

(b) Number of emergency admissions per year

Figure 4.9: Boxplots of outcomes hospitalisation and emergency admissions per year in the HF dataset.

Female and male patients have different comorbidities prevalence, including differences in the most common comorbidities and in the interaction between them. Male patients have in general a higher percentage of comorbidities than female patients. They also present different prevalence of medical appointments, with female patients having a higher number of medical appointments in most medical specialities. These differences do not seem to be reflected in the outcomes, as the distribution for hospitalisations and emergency admissions is quite similar for both genders.

4.2 Clustering Algorithm and Number of Clusters k

To understand which clustering algorithm was more suitable for the HF dataset, we used three different approaches, all taking into consideration the mixed-type nature of the data. The results from the different clustering algorithms applied to the HF dataset are summarised in Table 4.1. The combination

Table 4.1: Values obtained for clustering metrics Average Silhouette Score. Calinski-Harabasz Index and Davies-Bouldin Score for the different clustering algorithms, namely, Gower's Distance and Hierarchical Clustering, Factor Analysis of Mixed Data (FAMD) and Hierarchical Clustering, and FAMD and K-Means. For Silhouette Score and Calinski-Harabasz a higher value indicates a better performance, for Davies-Bouldin a lower value is best.

Clustering Algorithm	k	Average Silhouette Score	Calinski-Harabasz	Davies-Bouldin
Gower Distance + Ward Hierarchical Clustering	3	0.153	918.620	1.859
	4	0.155	910.162	1.751
	5	0.153	810.238	1.797
FAMD + Ward Hierarchical Clustering	3	0.080	221.906	2.267
	4	0.082	225.429	2.325
	5	0.075	227.625	2.332
FAMD + K-Means	3	0.073	217.615	2.374
	4	0.078	201.213	2.420
	5	0.068	191.439	2.537

of Gower's distance matrix and Ward's Hierarchical Agglomerative Clustering is the one that produced clusters with the highest scores in all clustering metrics analysed.

4.2.1 Choice of cluster number k

The choice of the number of clusters (k) was based on several factors. A minimum of 375 patients was set to promote stability, the metrics Average Silhouette Score, Calinski-Harabasz, and Davies-Bouldin were analysed, and difference in the resulting cluster characteristics was also evaluated. The metrics score for each value of k was considered by a majority vote, that is, the best value of k is the one that has the highest score in the majority of the metrics considered. Table 4.2 shows the clustering evaluation metrics for clusters with $k = [2, 12]$. The values were the best for $k = 2$. However, the resulting clusters included a cluster with a very high number of patients and a cluster with few patients (under 375). There was also no clear differentiation between patient groups. The second value of k with the best scores was $k = 4$. It was also observed that the resulting clusters had statistically different characteristics and were also different from a clinical point of view, which was evaluated together with the HF specialist. Hence, the value chosen for the analysis was $k = 4$.

4.3 Heart Failure Patient Subgroups

The clustering analysis resulted in four patient clusters, which were characterised in terms of comorbidities and demographic factors. Table 4.3 shows the detailed characterisation of each cluster and of the HF dataset. p-values presented in the table were computed by comparing the distribution of the variables among clusters, using the Kruskal-Wallis test for continuous variables and Chi-squared tests for categorical variables.

Patients from Cluster1 tend to be elder males. They show the highest number of International Classification of Diseases (ICD)-9 codes and number of chronic diseases. It is also the cluster with the highest prevalence of almost all diseases, having a high percentage of patients with all the diseases analysed.

Table 4.2: Values obtained for clustering metrics Silhouette Score. Calinski-Harabasz Index and Davies-Bouldin Score for Hierarchical Clustering with Gower's Distance using k=[2.10]. For Silhouette Score and Calinski-Harabasz a higher value indicates a better performance, for Davies-Bouldin a lower value is best.

Clusters	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
2	0.273	1147.505	1.494
3	0.153	918.620	1.859
4	**0.155**	**910.162**	**1.751**
5	0.153	810.238	1.797
6	0.147	746.957	1.739
7	0.141	673.593	1.983
8	0.127	615.439	2.005
9	0.125	571.963	2.041
10	0.130	538.409	1.939
11	0.132	512.288	1.961
12	0.140	491.961	1.914

The most common comorbidities are Anaemia (88.47%), CKD (74.52%), Hypertension (70.22%), and Atrial Fibrillation (55.64%). There is also a high prevalence of ICM (49.43%) and Diabetes (30.42%). Patients belonging to this cluster also have very high values of NT-proBNP, usually related to a more severe state of HF. The median values for Sodium and Urea are above the reference values, which is in agreement with the high percentage of CKD and the median value for Hemoglobin is below the reference values, which is in agreement with the high percentage of patients with Anaemia.

Patients from Cluster2 are mostly older women. The most common comorbidities in this cluster are Hypertension (84.53%), Atrial Fibrillation and Obesity (33.95%). Prevalence of Hypertension and Obesity is also higher in this cluster than in any other. Cluster2 patients present the highest medium value of BMI, which is related to the high percentage of patients with Obesity, and the lowest median value of NT-proBNP.

Cluster3 is the largest cluster (n=1231). Patients of Cluster3 are older men (59.46%) and women (40.54%) and have generally lower disease prevalence than the ones from Cluster2. The only disease where the prevalence is higher than in other clusters is Anaemia (99.35%). The median value of Hemoglobin is below the reference value in this cluster, which is in agreement with the high percentage of patients with Anaemia. The median value of NT-proBNP is the second highest when compared to other clusters.

Patients from Cluster4 are the youngest compared to other clusters and are predominantly female. These patients have the lowest number of diseases and the lowest prevalence of almost all diseases, except for Obesity (11.64%) and Cardiomyopathy (4.31%) where the prevalence is slightly higher than in Cluster3.

Figure 4.10 shows a tileplot of the clusters'. percentages of comorbidities that allows to easily find the most prevalent comorbidities in each cluster and compare the prevalence of diseases among clusters.

Table 4.3: Characterisation of the four resulting clusters obtained from the Hospital da Luz Lisboa HF dataset. Continuous variables are described as median (inter-quartile range) and categorical variables as %. p-values for the comparison of the characteristics across clusters.

Characteristics	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Dataset**	**p-value**
Number of Patients	789	866	1231	859	3745	-
Female, %	37.39	63.51	40.54	73.92	52.84	< 0.01
Male, %	62.61	36.49	59.46	26.08	47.16	< 0.01
Age, years	85.0(78.0-89.0)	81.0(73.0-87.0)	83.0(76.0-89.0)	76.0(59.5-85.0)	82.0(73.0-88.0)	< 0.01
BMI, kg/m^2	24.69(23.31-29.0)	26.14(24.16-31.09)	24.27(23.21-25.7)	24.83(23.44-27.06)	24.8(23.44-27.78)	< 0.01
Ischemic Cardiomyopathy, %	49.43	25.87	17.3	8.73	24.09	< 0.01
Cardiomyopathy, %	7.73	9.93	1.79	4.31	5.5	< 0.01
Hypertension, %	70.22	84.53	52.07	22.35	56.58	< 0.01
Diabetes, %	30.42	16.17	6.17	1.16	12.44	< 0.01
Atrial fibrillation, %	55.64	45.5	22.5	18.04	33.78	< 0.01
Transient Ischemic Attack, %	26.62	17.55	5.85	2.44	12.15	< 0.01
Valvular Disease, %	29.91	24.6	12.19	4.77	17.09	< 0.01
Chronic Kidney Disease, %	74.52	21.48	12.75	0.35	24.94	< 0.01
Anaemia, %	88.47	10.62	99.35	0.0	53.75	< 0.01
COPD, %	24.59	12.47	6.66	2.56	10.84	< 0.01
Obesity, %	25.86	33.95	6.99	11.64	18.26	< 0.01
Number of ICD-9 Codes	14.0(8.0-23.0)	10.0(6.0-16.0)	5.0(2.0-9.0)	4.0(2.0-7.0)	7.0(3.0-13.0)	< 0.01
Number of Chronic Diseases	8.0(6.0-10.0)	6.0(5.0-8.0)	4.0(2.0-5.0)	3.0(2.0-4.0)	5.0(3.0-7.0)	< 0.01
Sodium, mEq/L	140.0(137.0-143.0)	140.0(138.0-142.0)	139.56(137.0-142.0)	139.51(139.0-141.0)	139.57(138.0-142.0)	< 0.01
Urea, mg/dL	81.0(53.0-120.0)	51.0(39.0-68.0)	56.64(43.0-78.0)	44.94(35.78-55.67)	54.0(40.0-79.0)	< 0.01
Creatinin, mg/dL	1.68(1.22-2.36)	1.1(0.89-1.34)	1.23(0.98-1.52)	1.01(0.83-1.16)	1.17(0.93-1.56)	< 0.01
Hemoglobin, g/dL	10.7(9.3-11.9)	13.3(12.1-14.4)	11.0(9.8-12.3)	13.4(12.47-14.2)	12.19(10.6-13.4)	< 0.01
Red Cell Distribution Width, %	15.7(14.4-17.3)	14.39(13.4-15.3)	14.89(14.3-16.3)	14.3(13.4-14.75)	14.69(13.8-15.9)	< 0.01
Platelet count, x10^3/L	209.0(152.0-264.0)	214.0(172.0-253.81)	225.0(182.0-276.0)	238.0(193.5-264.0)	223.0(176.0-263.08)	< 0.01
NT-proBNP, pg/ml	4873.0(1648.0-13342.0)	1440.5(402.5-4083.19)	3817.0(1455.0-6973.7)	2173.0(531.5-4275.74)	2800.1(942.0-6255.0)	< 0.01
Number of Consultations/year	4.5(1.0-12.5)	3.5(1.0-8.5)	01.0(0.0-3.5)	1.0(0.0-3.5)	2.0(0.5-6.0)	< 0.01

It is possible to confirm that Cluster1 has the highest percentage of almost all diseases, whereas Cluster4 can be seen as the low-burden cluster, considering the low percentage of other comorbidities. Cluster2 can be characterised by the high prevalence of Hypertension, Atrial Fibrillation and Obesity, and Cluster3 is the cluster with the highest percentage of patients with Anaemia.

Figure 4.11 shows the graph representation of each cluster's comorbidities. The representation can provide insights into cluster complexity. It is possible to see that Cluster1 has the highest number of nodes and edges and the graph of Cluster4 has the lowest, which indicates that patients from Cluster1 have a higher complexity than patients from Cluster4. Figure 4.11 also illustrates the most common comorbidity associations in each cluster. Cluster1 has a high number of associations and a high number of strong associations between diseases. The fact that all nodes are of similar size means that every disease co-occurs at least once with almost all other diseases. The width of the edges is what allows us to understand which of these co-occurrences are more common. In Cluster1, these are Chronic Kidney Disease and Anaemia, and Hypertension and Anaemia. In Cluster2 there is a high number of patients with Obesity and Atrial Fibrillation, and Hypertension and Atrial Fibrillation. Cluster3 has a

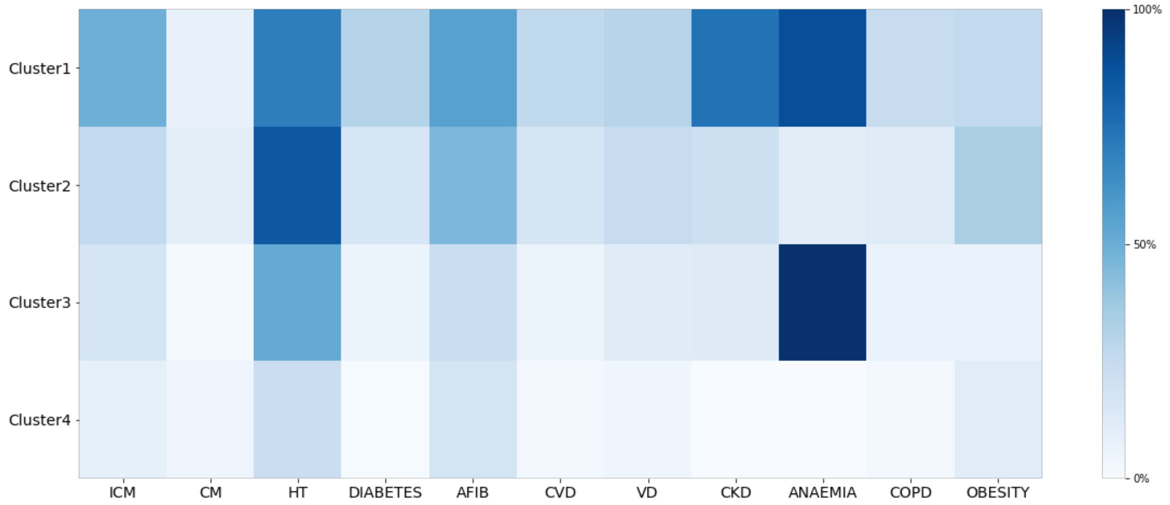
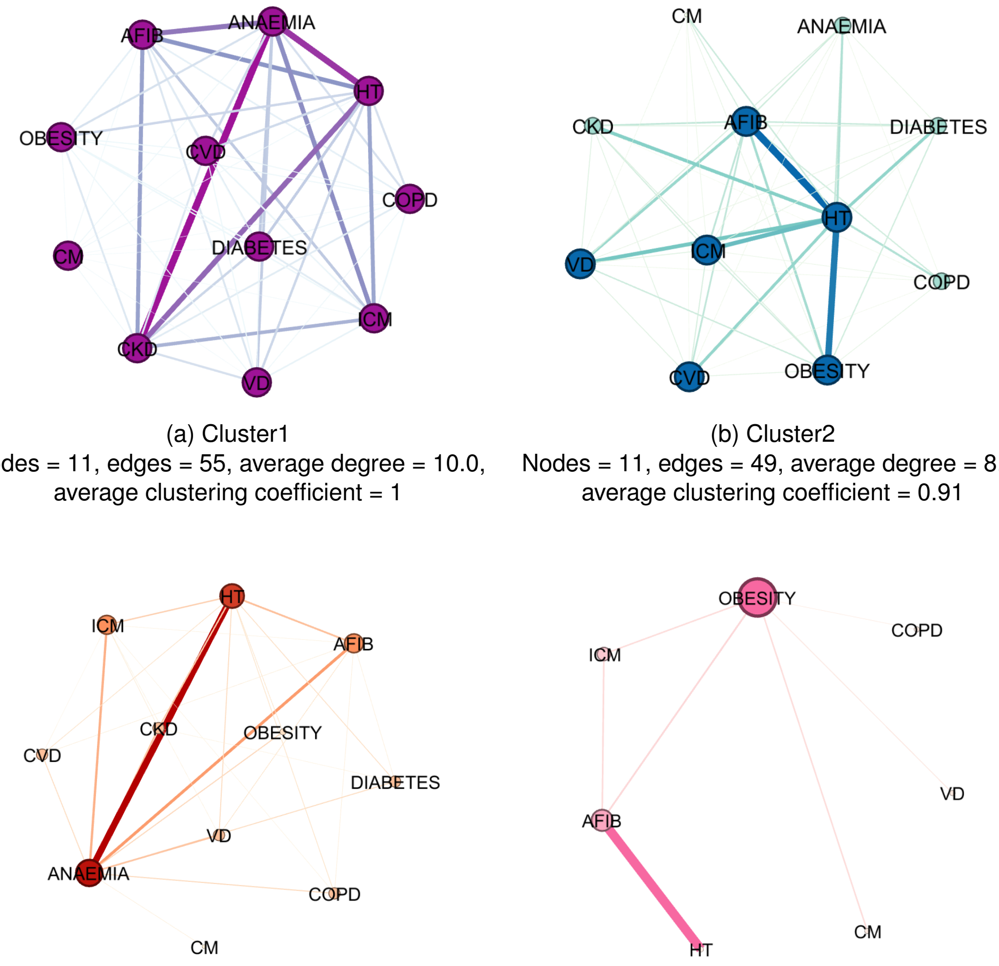


Figure 4.10: Cluster-specific percentages of comorbidities. A darker colour indicates a higher percentage of the comorbidity in the cluster. ICM-Ischaemic Cardiomyopathy, HT-Hypertension, AFIB-Atrial Fibrillation, CVD-Cerebrovascular Disease, VD-Valvular Disease, CKD-Chronic Kidney Disease, COPD-Chronic Obstructive Pulmonary Disease

high number of patients with Hypertension and Anaemia, with some of these patients also having Atrial Fibrillation, Ischaemic Cardiomyopathy, and Valvular Disease. Looking at the graph from Cluster4, the most common association of diseases is Hypertension and Atrial Fibrillation. Obesity is also connected to several other diseases in this cluster, but with a lower co-occurrence. The metrics average degree and average clustering coefficient are also helpful to quantify the clusters' complexity. The average degree indicates the average number of other diseases that are connected to one disease. We can observe that in Cluster1 the average degree is 10, which means all diseases are connected to each other. Cluster2 also has a high degree, 8.9, whereas Cluster3 and Cluster4 have much lower degrees, 5 and 2, respectively. The average clustering coefficient is a measure of density that indicates the degree to which nodes in a graph tend to cluster together. In this case, it can be interpreted as a measure of the tendency of diseases to co-occur. Also for this metric Cluster1 has the highest value (1) and Cluster4 has the lowest (0.48), which suggests patients from Cluster1 have a higher order co-occurrence of diseases than patients from Cluster4.

In terms of comorbidities, the identified clusters present similarities with the results reported in a study by Gulea et al. (2021). The study applied model-based clustering to 12 comorbidities of a cohort of HF patients and identified five clusters. The clusters were characterised by a different combination of comorbidities and socio-demographic factors, and named accordingly: low-burden, metabolic-vascular, Ischaemic, anaemic, and metabolic. The study found that patients in the metabolic-vascular cluster had the highest percentage of comorbidities and worst prognosis and that the patients in the low-burden cluster had the lowest percentage of comorbidities and best prognosis. In the results presented, it is also possible to identify a metabolic-vascular cluster, corresponding to Cluster1, an anaemic cluster, corresponding to Cluster3, and a low-burden cluster, corresponding to Cluster4. Cluster2, which can be identified as an obesity cluster, presents some similarities with the metabolic cluster found by Gulea et al. (2021), however, it lacks the high prevalence of Diabetes to be considered a metabolic cluster.



(a) Cluster1
Nodes = 11, edges = 55, average degree = 10.0,
average clustering coefficient = 1

(b) Cluster2
Nodes = 11, edges = 49, average degree = 8.9,
average clustering coefficient = 0.91

(c) Cluster3
Nodes = 11, edges = 28, average degree = 5.0,
average clustering coefficient = 0.89

(d) Cluster4
Nodes = 7, edges = 7, average degree = 2.0,
average clustering coefficient = 0.48

Figure 4.11: Graph representation of comorbidities present in the obtained clusters. A node represents a disease and its size is proportional to the node degree. An edge represents a co-occurrence of two diseases and its width is proportional to the prevalence of the co-occurrence in the cluster. ICM-Ischaemic Cardiomyopathy, HT-Hypertension, AFIB-Atrial Fibrillation, CVD-Cerebrovascular Disease, VD-Valvular Disease, CKD-Chronic Kidney Disease, COPD-Chronic Obstructive Pulmonary Disease

4.3.1 Prescriptions

Data on medical prescriptions were available for 2838 patients (75.78% of the overall study population). In terms of the distribution of these patients in each cluster, in Cluster1 data were available for 691 patients (87.58%), in Cluster2, 731 patients (84.41%), in Cluster3, 849 patients (68.97%) and in Cluster4, 567 (66.00%).

Table 4.4 shows the results obtained for the mean number of prescriptions and percentages of specific medication groups in each cluster. The number of different medications is relative to the period analysed for medical prescriptions (9 years and 8 months). Percentages of medication groups represent the percentage of patients that had at least one prescription of the relevant medication group during the period analysed.

The average number of prescriptions per year for the entire population, during the period analysed, is approximately 4. Cluster1 has the highest number of prescriptions, with an average of 7 prescriptions per year. Cluster2 is also above average with approximately 4, while Cluster3 and Cluster4 are below the dataset average, with Cluster3 having a higher average number of prescriptions than Cluster4.

Medication for HF treatment includes Angiotensin-Converting-Enzyme Inhibitors (ACEi)s\Angiotensin Receptor Blockers (ARB)s, Beta-blockers, Diuretics, and Digoxins (European Society of Cardiology, 2016). All of these groups are amongst the most prevalent medication groups in the dataset. Besides these medication groups, patients also have a high prevalence of Anticoagulants, Antiplatelets, Statins and Bronchodilators.

Cluster1 can be considered the cluster with the most severe stage of HF and thus it is expected that a high percentage of patients have prescriptions directly related to HF and also related to their comorbidities. The most common medication groups in each cluster are Anticoagulants (42.98%), ACEi\ARB (40.96 %), Statins (39.94%) and Beta-blockers (35.31%). There is also a high percentage of patients with prescriptions for Bronchodilators (32.27%), which are used for the treatment of COPD, and Diuretics (29.52%), commonly used to treat Hypertension or edema, often consequences of HF or CKD. In Cluster2 the most common medication group is Anticoagulants (37.76%), followed by Beta-blockers (35.16%), Statins (33.52%) and ACEi\ARB (28.32%). In this cluster, the most common comorbidities were Hypertension, Obesity, and Atrial Fibrillation, which are in agreement with the most common medication groups found. Patients from Cluster3 and Cluster4 were found to have a lower percentage of all medication groups when compared to Cluster1 and Cluster2, with Cluster4 having the lowest percentage of all medication groups except for Levothyroxine, Digoxin and Sodium-Glucose Transport Proteins (SGLT)2i. It is possible to observe that only 16.37% of patients in Cluster4 have been prescribed Hematinic factors while 99.35% of patients in this cluster were found to have Anaemia. Hematinic factors are often used to treat ferropenic Anaemia (Anaemia caused by iron deficiency). This is possibly due to an underdiagnosis of Anaemia or to Anaemia being due to other factors than iron deficiency.

4.3.2 Medical Appointments

Data on medical appointments were available for 2705 patients (76.23% of the overall study population). In terms of the distribution of these patients in each cluster, in Cluster1 data were available for 618 patients (78.32%), in Cluster2, 707 patients (81.63%), in Cluster3, 776 patients (63.11%) and in Cluster4, 603 (70.19%). It is important to note that the medical appointments data collected may not represent the whole reality of this dataset, as patients can also attend medical appointments in other healthcare facilities.

Figure 4.12 presents the distribution of medical appointment attendance by speciality and by cluster. Cardiology is the medical speciality with the highest percentage of patients for all clusters, followed by Internal Medicine. From the cluster profiles traced so far, Cluster1 is, overall, the cluster with the highest complexity, presenting a higher percentage of comorbidities and medications, and this complexity is also translated into a higher percentage of medical appointments. Cluster2 is the cluster with the second highest number of medical appointments' percentage, followed by Cluster3 in the majority of medical specialities. Cluster4 stands out for the high percentage of patients attending Gastroenterol-

Table 4.4: Number of medication and medication groups prevalence per cluster and in the entire dataset. p-values for the comparison of the characteristics across clusters. ACEi - angiotensin–converting enzyme inhibitor; MRA - Aldosterone receptor antagonists; DPP4i - Dipeptidyl peptidase-4 inhibitor; ARB - angiotensin receptor blockers; ARNi - Angiotensin Receptor-Neprilysin Inhibitors; SGLT2i - Sodium-glucose cotransporter 2 inhibitors; GLP-1-Glucagon-like peptide-1.

Medications	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**	**Dataset**	**p-values**
Patients with medication data	87.58	84.41	68.97	66.0	75.78	-
Avg prescriptions/year	6.6	4.15	2.53	2.00	3.83	-
Anticoagulants	42.98	37.76	26.5	19.58	32.03	<0.01
Statins	39.94	33.52	21.91	18.34	28.58	<0.01
Beta-Blockers	35.31	35.16	22.38	17.28	27.8	<0.01
ACEi \ARBs	40.96	28.32	21.29	17.11	27.20	<0.01
Antiplatelets	34.88	24.62	21.08	10.76	23.29	<0.01
Inhalers Bronchodilator	32.27	23.53	16.49	13.76	21.6	<0.01
Diuretics	29.52	26.54	15.19	10.58	20.68	<0.01
Hematinic factors	27.79	15.18	16.37	8.47	17.27	<0.01
Anticholinergics	23.59	15.73	11.9	8.64	15.08	<0.01
MRA	16.06	14.5	12.25	7.58	12.83	<0.01
DPP4i	19.83	9.58	6.24	2.29	9.62	
Antiarrhythmics	11.0	9.99	6.24	5.29	8.17	<0.01
Levothyroxine	9.84	7.11	5.65	6.35	7.19	<0.05
ARN	9.99	8.34	5.77	3.53	7.01	<0.01
Metformin	10.85	6.7	5.54	3.7	6.77	<0.01
Insulin	13.75	4.51	3.77	0.88	5.81	<0.01
Digoxin	6.8	7.39	2.83	4.41	5.29	
Sulfonylureas	8.83	2.46	2.59	2.47	4.05	
Ivabradine	3.62	2.6	2.94	1.06	2.64	<0.05
SGLT2i	3.91	3.15	0.82	1.06	2.22	<0.01
Inhaled Corticosteroids	2.89	2.46	0.71	0.71	1.69	<0.01
GLP-1 antagonists	1.16	1.23	0.12	0.35	0.7	<0.05

ogy appointments. This cluster also has the highest percentage of Obstetrics-Gynecology (OB-GYN) consults, which is in agreement with the demographic profile of the cluster (younger women).

The graph representation of the clusters medical appointments' transitions can be seen in Figure 4.13. As explained previously, the size of the node is indicative of the number of connections to other nodes, the node degree, and the width of the edges is proportional to the number of patients that underwent that specific transition. Observing the graphs of the different clusters and the number of nodes, edges and average degree indicated in the caption, it is possible to understand which clusters have higher complexity in terms of medical appointments. Cluster1 and Cluster2 graphs have a higher number of different transitions and also have more patients performing each transition, which is in line with the higher percentage of medical appointments seen before. Comparing the graph's from Cluster3 and Cluster4 it is possible to note that while Cluster4 has a higher number of nodes and edges, Cluster3 has a higher average degree. This indicates that patients from Cluster4 have a higher variety of medical appointments and transitions but each node is often only linked to one other node, patients from Cluster3 have more recurrent transitions and medical specialities are more interconnected to each other.

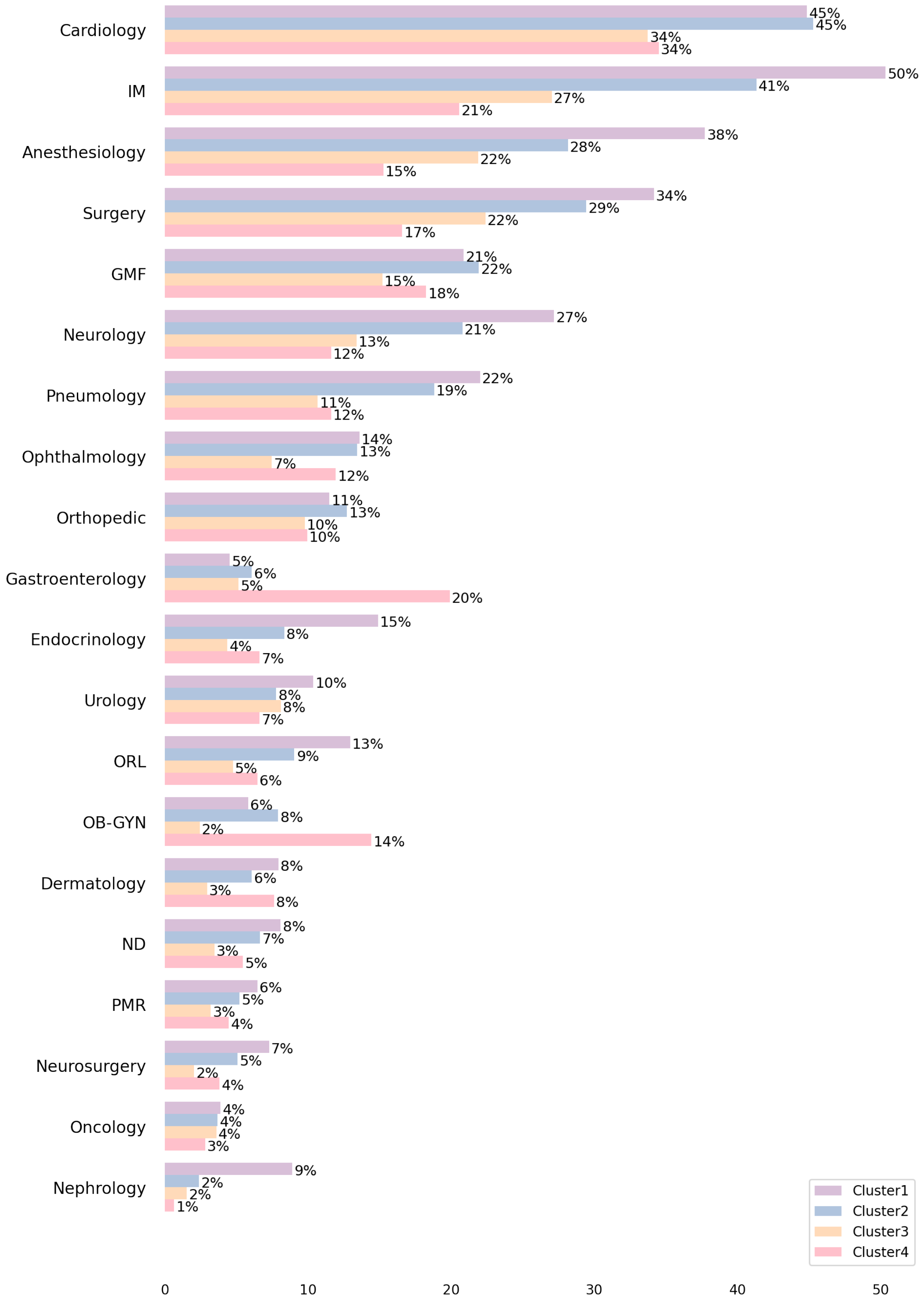
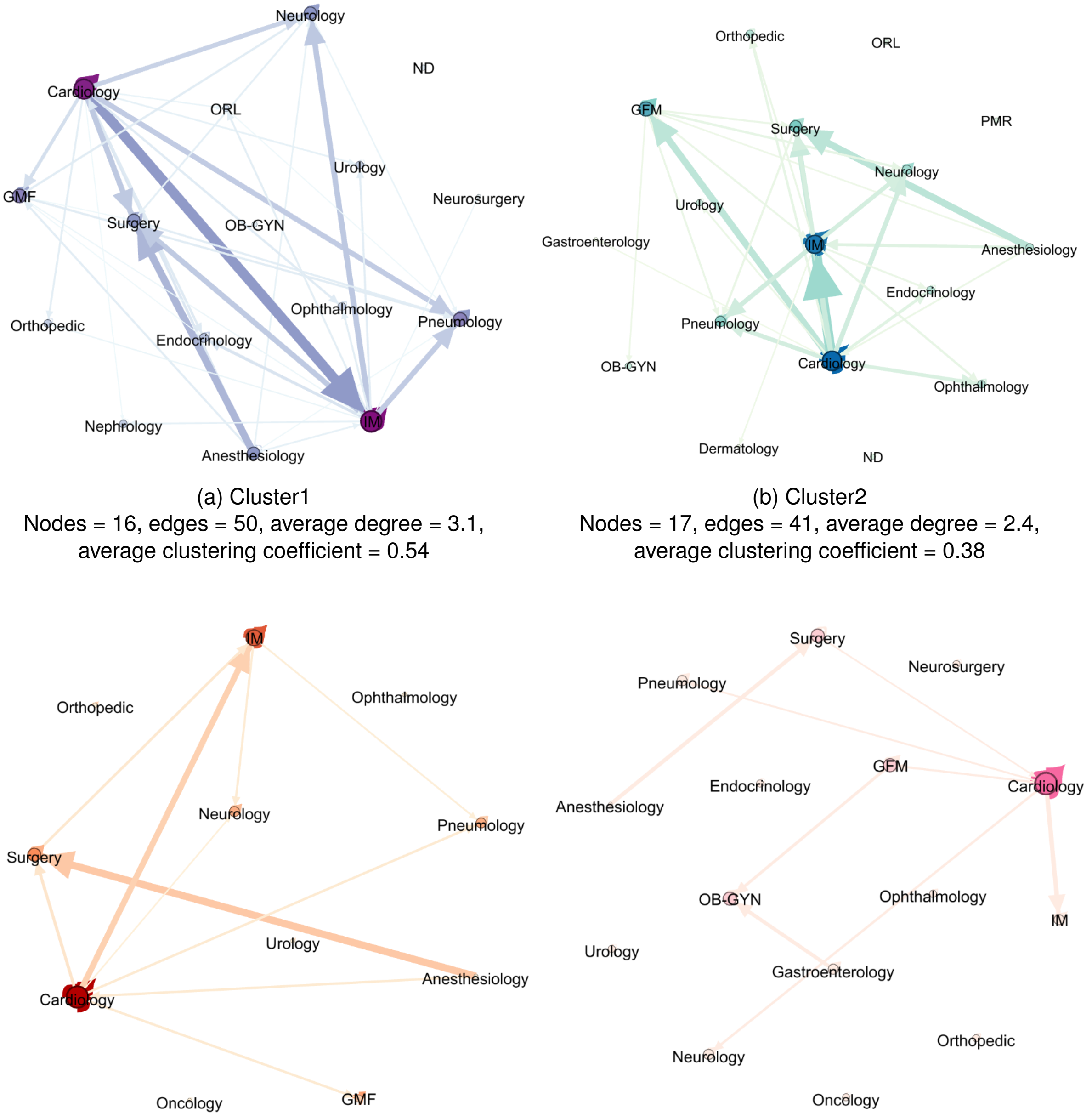


Figure 4.12: Distribution of medical appointments in each cluster. ND - Nutrition and Dietetic, PMR-Physical Medicine and Rehabilitation, GFM - General and Family Medicine, IM - Internal Medicine.

Despite the limitations indicated above, the analysis of the medical appointments provides insights into another domain of the clusters, improving their characterisation. It is also useful to understand the differences in the use of healthcare resources among clusters. The results for medical appointments follow the cluster profiles from previous sections, with Cluster1 having a higher healthcare utilisation and the highest percentage of medical appointments and transitions, followed by Cluster2, and Cluster3 and Cluster4 presenting a lower percentage of medical appointments and transitions.



(a) Cluster1
Nodes = 16, edges = 50, average degree = 3.1,
average clustering coefficient = 0.54

(b) Cluster2
Nodes = 17, edges = 41, average degree = 2.4,
average clustering coefficient = 0.38

(c) Cluster3
Nodes = 11, edges = 20, average degree = 1.8,
average clustering coefficient = 0.36

(d) Cluster4
Nodes = 15, edges = 22, average degree = 1.4,
average clustering coefficient = 0.17

Figure 4.13: Graph representation of medical appointments and their transisitons in each cluster. A node represents a medical appointment speciality and its size is proportional to the node degree. An edge represents a transition from one medical appointment to another and and its width is proportional to the prevalence of that transition in each cluster. Medical appointments by gender. ND- Nutrition and Dietetic, PMR- Physical Medicine and Rehabilitation, GFM - General and Family Medicine, IM - Internal Medicine.

Table 4.5: Characterisations of outcomes related variables per cluster and in the entire dataset. Continuous variables are described as median (inter-quartile range) and categorical variables as %. p-values for the comparison of the characteristics across clusters.

Characteristics	Cluster1	Cluster2	Cluster3	Cluster4	Dataset	p-value
Number of Hospitalisations/year	0.2(0.1-0.4)	0.1(0.0-0.2)	0.1(0.0-0.2)	0.0(0.0-0.1)	0.1(0.0-0.2)	<0.05
Hospitalisations within 1 year of HF diagnosis, %	35.23	23.78	27.86	21.18	15.25	<0.01
Hospitalisations within time period analysed, %	86.06	68.13	59.46	32.6	60.91	<0.01
Number of Emergency admissions/year	0.6(0.2-1.2)	0.4(0.1-0.8)	0.2(0.1-0.5)	0.1(0.1-0.4)	0.3(0.1-0.7)	<0.05
Emergency admissions within 1 year of HF diagnosis, %	52.47	44.34	39.48	36.20	28.44	<0.01
Emergency admissions within time period analysed, %	94.55	88.45	79.45	75.09	83.71	<0.01
Deceased, %	29.02	12.7	18.28	5.82	16.4	<0.01

Table 4.6: Odds ratio between each cluster and other clusters for the outcomes mortality, hospitalisation and emergency. An odds ratio of 1.8 in the outcome Emergency for Cluster2 vs others, for example, means that patients in Cluster2 are 1.8 times more likely have an emergency admission than patients in other clusters.

Outcome	Clusters	Odds ratio	p-value
Death	Cluster1	2.73	<0.05
	Cluster2	0.69	<0.01
	Cluster3	1.22	<0.05
	Cluster4	0.35	<0.01
Hospitalisation	Cluster1	4.96	<0.1
	Cluster2	1.13	= 0.01
	Cluster3	0.74	<0.01
	Cluster4	0.14	<0.01
Emergency	Cluster1	3.18	<0.01
	Cluster2	1.80	<0.01
	Cluster3	0.62	<0.01
	Cluster4	0.4	<0.01

4.3.3 Outcomes

The outcomes of hospitalisation, emergency admission, and mortality were also analysed to understand if the clustering resulted in groups that had different risk associations with the outcomes considered. One important consideration for this analysis is that only events occurring during the period analysed in HLL are considered, and so it is possible that some patients experienced these events outside that time window or in other healthcare facilities.

The percentage of patients in the dataset with at least one hospitalisation is 60.91% and with at least one emergency admission is 83.71% (see Table 4.5). Cluster1 is the highest complexity cluster, which also translates into a higher percentage of patients that experienced hospitalisations and emergency admissions, and deceased patients. It is also the cluster with the highest number of hospitalisations and emergency admissions per year. Cluster2 has the second highest percentage of number of hospitalisations higher than one, while Cluster3 has the second highest percentage of deceased patients. Cluster4, the low-burden cluster, has the lowest percentages for all outcomes considered.

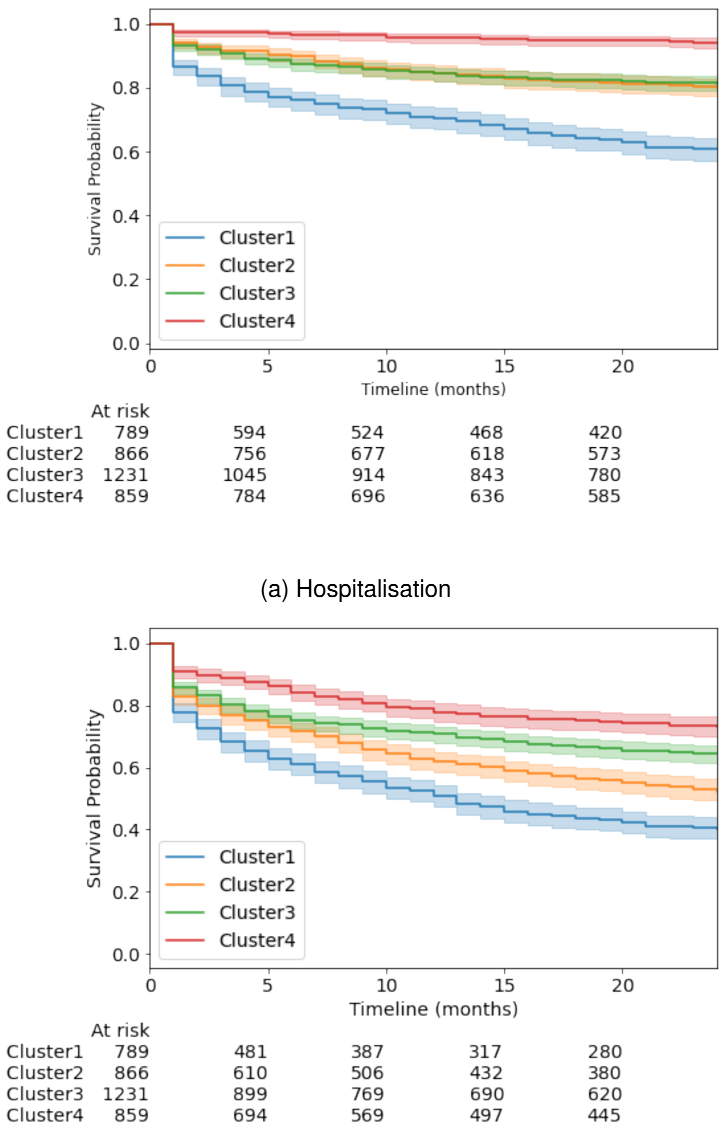
For each outcome the Odds Ratio (OR) between the patients in each cluster and all the patients outside that cluster was computed, the results are presented in 4.6. The goal of OR is to determine the association between two variables, in this case, between each cluster and each outcome.

Cluster1 presents the highest values of OR for all outcomes considered, with 2.73 for the outcome mortality, 4.96 for the outcome hospitalisation and 3.18 for the outcome emergency admission. These values mean that patients from Cluster1 are 2.73 more likely to be dead, 4.96 more likely to be hospitalised and 3.18 more likely to have an emergency admission than patients from outside the cluster. These results are in agreement with the cluster profile defined so far, as Cluster1 is the cluster with the eldest patients and the highest comorbidity rates. Inversely, Cluster4 is the cluster with the lowest OR for all outcomes, with 0.35 for the outcome mortality, 0.74 for the outcome hospitalisation and 0.4 for the outcome emergency admission. OR lower than 1 indicate that patients are less likely to experience the outcomes when compared to other patients. Cluster2 and Cluster3 both have OR higher and lower than 1, depending on the outcome considered. Patients from Cluster2 have an OR of 0.69 for the outcome mortality, indicating that they have a lower risk for this outcome, and OR of 1.13 and 1.80 for the outcomes hospitalisation and emergency, respectively. indicating a higher risk for these outcomes. For patients from Cluster3, the opposite is true, since the cluster presents OR of 1.22 for the outcome mortality and OR of 0.74 and 0.14 for the outcomes hospitalisation and emergency, respectively.

4.3.4 Cluster Survival Analysis

Until this point, the clusters have been characterised in the domain of demographics, comorbidities, laboratory values, medical prescriptions, consultations and outcomes prevalence. It was then necessary to understand if there was an association between belonging to a cluster with a certain profile and the probability of experiencing a given clinical outcome. To do so, a survival analysis was conducted for the outcomes of hospitalisation and emergency for all patients in the dataset. Survival analysis provides methods to determine how one group changes relative to another, in this case, how being in one cluster relative to being in another influences the outcome. This analysis was not possible to conduct for the mortality outcome due to the lack of longitudinal information.

Figure 4.14 shows univariate Kaplan-Meier curves for the outcomes of hospitalisation and emergency admission, stratified by clusters, for a time period of 2 years post HF diagnosis. Kaplan-Meier curves represent the survival probability of a population over time. In this case, each cluster represents a different population and the survival probability is represents not experiencing a hospitalisation or an emergency admission. At t_0 all clusters have a survival probability of 1 and as time moves forward the survival probability goes down as a function of the number of patients experiencing the outcome in each cluster. Cluster1, as could be expected from the previous characterisation, appears as the highest risk cluster, with the lowest survival probability at all times. Contrarily, Cluster4 shows the highest survival probability during the period analysed. Cluster2 and Cluster3 have a similar survival function for the outcome hospitalisation. However, for the outcome of emergency admission Cluster3 has a lower survival probability than Cluster3. It is also interesting to note that survival probability for the outcome emergency admission is, for all clusters, lower than for the outcome hospitalisation, suggesting that HF patients are more likely to have emergency admissions rather than being hospitalised. Differences between groups were tested using the log-rank test and the obtained p-value was $<$0.01 for both outcomes.



(a) Hospitalisation

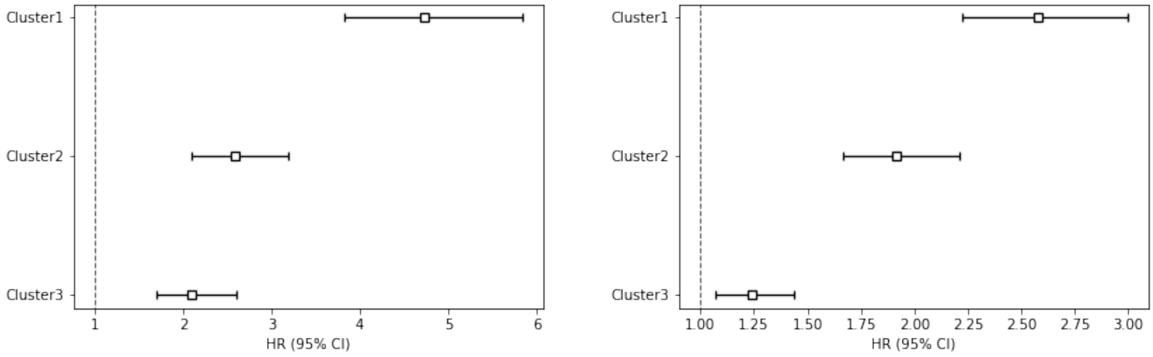
(b) Emergency Admission

Figure 4.14: Kaplan-Meier survival curves for the outcomes Hospitalisation and Emergency admission for each cluster (within 2-years after HF diagnosis).

Besides Kaplan-Meier curves, Cox proportional hazard models were also computed to study how belonging to a certain cluster changes the rate of experiencing an outcome, with the possibility of taking into account other variables. One unadjusted and two multivariable adjusted Cox proportional hazards were computed (results are displayed in Table 4.7). Model1 is unadjusted, Model2 is adjusted for baseline covariates Age and Gender, and Model 3 is adjusted for baseline covariates Age and Gender, and for NT-proBNP, often used as a risk marker for HF. Figure 4.15 presents the graphical representation for Model3. The Hazard Ratio (HR) are computed in relation to the lowest risk cluster, Cluster4. All clusters show a higher risk for the outcomes of hospitalisation and emergency admission. This risk is the highest for Cluster1 and the lowest for Cluster3. In Model1, HR for hospitalisation ranged from 5.86 (4.80 - 7.15) for Cluster1 to 2.43 (1.98 - 2.98) for Cluster3, when compared with Cluster4. For the outcome hospi-

Table 4.7: Risk of clinical events hospitalisation and emergency admission compared with Cluster4 (lowest risk). Hazard ratios and 95% confidence intervals computed using Cox Regression. Model1 unadjusted, Model3 adjusted for Age and Gender, Model3 adjusted for Age, Gender and NT-proBNP.

	Cluster1	Cluster2	Cluster3	Cluster4	p-value
Model1, HR (95% CI)					
Hospitalisation	5.86 (4.80 - 7.15)	2.82 (2.29 - 3.48)	2.43 (1.98 - 2.98)	1	$<$0.05
Emergency Admission	2.73 (2.38 - 3.14)	2.00 (0.84 - 2.31)	1.29 (1.13 - 1.49)	1	$<$0.05
Model2, HR (95% CI)					
Hospitalisation	4.90 (3.97 - 6.05)	2.57 (2.08 - 3.18)	2.10 (1.70 - 2.60)	1	$<$0.05
Emergency Admission	2.60 (2.24 - 3.02)	1.92 (1.66 - 2.21)	1.24 - (1.07 - 1.44)	1	$<$0.005
Model3, HR (95% CI)					
Hospitalisation	4.73 (3.83 - 5.84)	2.58 (2.09 - 3.19)	2.10 (1.70 - 2.60)	1	$<$0.05
Emergency Admission	2.58 (2.22 - 2.99)	1.92 (1.66 - 2.21)	1.24 (1.07 - 1.44)	1	$<$0.05



(a) Hospitalisation

(b) Emergency

Figure 4.15: Risk of clinical events hospitalisation and emergency admission compared with Cluster4 (lowest risk). Symbols represent hazard ratios and 95% confidence intervals computed using Cox Regression adjusted for Age, Gender and NT-proBNP (Model3).

talisation, HR ranged from 2.73 (2.38 - 3.14) for Cluster1 to 2.00 (0.84 - 2.31) and 1.29 (1.13 - 1.49) for Cluster3. Despite having slightly lower values, differences in the HR of both outcomes remained significant when adjusting for Age and Gender, in Model2, and also when adjusting for NT-proBNP, in Model3.

The results obtained in this section are in agreement with the cluster profiles obtained from previous sections. Cluster1 has the highest risk profile, with older patients having a high number of comorbidities. Cluster2, a slightly younger cluster, with a high prevalence of AFIB, Obesity and HT, has the second highest risk for hospitalisation and emergency admissions. Cluster3 patients, with a high prevalence of Anaemia, show moderate risk for the outcomes considered. The lowest risk cluster is Cluster4, whose patients are younger and have the least prevalence of comorbidities. The results obtained show that different clusters are associated with different levels of risk for the outcomes considered, and thus cluster membership can be used for risk stratification.

In terms of risk association and cluster characteristics, the clusters presented similarities with a study by Ahmad et al. (2014). The study identified 4 clusters in a HF population, which were associated with different risk levels, having found a higher risk cluster and lower risk cluster. The higher risk cluster and lower risk cluster have some points in common with Cluster1 and Cluster4 identified in this dissertation. The higher risk cluster from Ahmad et al. (2014) and Cluster1 from our study had the eldest patients and were mainly male, with a high prevalence of HT and ICM, patients from the lower risk cluster and Cluster4 were mainly females and had a lower prevalence of comorbidities than all other patients.

4.4 Summary

This chapter presented the evaluation of the HF patient clusters obtained from applying the proposed workflow to Electronic Health Record (EHR) from HLL. The HF dataset from HLL was found to represent an elder population, balanced in terms of gender and prone to have a high number of chronic conditions, with more than 50% of the population having over five chronic diseases. The most common comorbidities were Hypertension, Anaemia and Atrial Fibrillation. Male and female patients presented differences in terms of comorbidities and medical appointments, with male patients having a higher prevalence of comorbidities and female patients having a higher number of medical appointments.

Using Ward's Hierarchical Clustering together with Gower's distance, four HF patient subgroups were obtained, that presented significant differences in the domains analysed. Cluster1 (the metabolic-vascular cluster) was the highest risk cluster, with the eldest patients and the highest number and prevalence of comorbidities. Patients from this cluster have the highest values for NT-proBNP and show a 4.73 and 2.58 HR for the outcomes of hospitalisation and emergency admission, respectively, when compared to patients from Cluster4. Patients from Cluster2 (the obesity cluster) were predominately females with a tendency for Obesity, Hypertension and Atrial Fibrillation. These patients are at the second highest risk for hospitalisation and emergency admission outcomes. Cluster3 (the anaemic cluster) has the highest percentage of patients with Anaemia. These patients have the second highest values for NT-proBNP and the second highest mortality rate. Patients from Cluster4 (the low-burden cluster) are mainly female, have the lowest risk for the outcomes considered, the lowest number of diseases and are the youngest when compared to other clusters.

It was possible to compare the clusters obtained with previous studies and find similarities. The phenotypes obtained for the clusters present similarities with a previous study by Gulea et al. (2021), where a metabolic cluster, an anaemic cluster and a low-burden cluster were also found. The clusters were also compared with a study by Ahmad et al. (2014), who also identified a high risk and a lower risk cluster.

Chapter 5

Conclusions

This dissertation presented a workflow to identify and phenotype subgroups of HF patients with multimorbidity, using real-world EHRs data from HLL.

This project was motivated by the need to study patients with multimorbidity, with a special interest in HF patients. Patients with multimorbidity represent a heavy burden for the healthcare system and are among the most difficult to treat patients. However, there is still a lack of knowledge about the most common disease interactions and risk factors in multimorbidity, as well as the efficacy and cost-effectiveness of different therapies for these patients. This is particularly important in complex and heterogeneous diseases such as HF, where an improved classification and characterisation of the disease can aid healthcare professionals in defining treatment strategies and designing clinical trials.

The developed workflow defines a method for selecting HF patients with multimorbidity, using ICD-9 codes and specific keywords from unstructured data, followed by preprocessing, including missing data imputation and characterisation of the HF dataset, patient clustering and a statistical and survival analysis of the resulting clusters.

The clustering algorithm and number of clusters k that performed best according to the scoring system defined were Ward's Hierarchical Clustering together with Gower's distance and $k = 4$. The four resulting clusters presented significant differences in clinical and demographic characteristics, along with different prevalence of comorbidities, and we were able to trace a clinical profile for each cluster. Medical appointments and medical prescriptions also presented differences between the patient subgroups, contributing to the clusters' profiles.

A survival analysis showed that each HF subgroup is associated with different levels of risk for the outcomes of hospitalisation and emergency admission. The results from the survival analysis were in agreement with the profile traced for each cluster, clusters with a higher complexity (higher number of comorbidities, medical prescriptions and medical appointments) had a worse prognosis and patients with a lower complexity had a better prognosis.

Comparisons between EHRs phenotyping studies can be challenging since the baseline populations do not have the same characteristics, which will also influence the characteristics of the clusters. However, considering the main characteristics of each cluster, such as the most prevalent comorbidities and their association with risk, we were able to compare the results obtained with previous studies. Ahmad et al. (2014) identified 4 clusters in a HF population, which were associated with different risk levels, having found a higher risk cluster and lower risk cluster that had similar characteristics to Cluster1 (high risk) and Cluster4 (low risk) identified in this dissertation. The phenotypes obtained for the clusters also presented similarities with a previous study by Gulea et al. (2021), where a metabolic cluster, an anaemic cluster and a low-burden cluster were also found.

The use of graphs as a representation of the cluster's comorbidities and medical appointment provided not only a visualisation tool but also metrics, such as the node degree and the clustering coefficient, that helped to quantify and compare the clusters complexity and higher order interactions between diseases.

This methodology is a first step in understanding the complexity of the HF subgroups, and may give suggestions for future research focus areas. For example, if the low complexity of Cluster4 is considered as this being the beginning ofHF related problems, we could study how to tailor surveillance and treatment to ensure patients from this cluster do not evolve to the other more complex clusters.

It is important that these results are always interpreted and validated from a clinical perspective. Moreover, creating a common vocabulary between scientists and healthcare professionals is fundamental to be able to implement changes in clinical decision support systems.

HF is a highly heterogeneous syndrome. In this study, it was possible to observe this with real patient data containing information from different sources, as clinical data, medical appointments, hospitalisations and emergency admissions, among others. This work showed that it is possible to obtain meaningful patient subgroup phenotypes using EHRs data available in a hospital (HLL). This in turn suggests that the proposed workflow can be a useful tool to HF specialists, such as the professional who followed this work. We hope this approach is able to contribute to improving the classification of this heterogeneous clinical syndrome, treatment strategies and the future design of HF clinical trials.

5.1 Limitations

The results of this work also present some limitations, mainly related to the characteristics of real-world health data and to the unavailability of specific information, such as the Ejection Fraction (EF) of patients.

The use of real-world data in the healthcare field, as in this thesis, is usually associated with several challenges. Rudrapatna and Butte (2020) review some of the most important challenges and limitations, highlighting data quality and data heterogeneity, bias and frequent incompleteness. In this study, these limitations are reflected in the identification of HF patients in the HLL database and of their comorbidities, which was done using ICD-9 codes and searching clinical text for specific keywords. It is possible that some coding errors exist, as for example not coding or miscoding a certain disease. Searching clinical text also presents a challenge due to the use of very specific terminology, abbreviations and misspellings, that can make the identification of diseases harder.

Regarding the classification of the HF syndrome, the parameter EF is determinant, and patients with Heart Failure with Reduced Ejection Fraction (HFrEF) and Heart Failure with Preserved Ejection Fraction (HFpEF) often present differences in the disease manifestation and response to treatment. Unfortunately, data from echocardiograms were not available for this work, preventing the inclusion of this information in the clustering stage. There was also no information regarding the severity of HF, which could provide helpful insights when evaluating differences between clusters.

Another limitation is related to the fact that data for appointments and medical prescriptions were not available for all the patients in the study. Moreover, it is also possible that the data for these domains and for the outcomes do not reflect the whole condition of the patient, as we only take into account prescriptions, appointments and outcomes taking place at HLL while it is possible that patients also make use of other healthcare facilities.

5.2 Future Work

Regarding the clustering and the characterisation of the clusters, future studies could benefit from integrating the EF parameter, either as a feature used for clustering or to analyse if the patient subgroups resulting from the clustering have similar EF values. It could also be interesting to study what type of interventions and device implementations, such as pacemakers, the patients have gone through.

With respect to the survival analysis, it would be interesting to consider only the events related to HF in addition to all-cause events, as for example in Shah et al. (2014). By doing so, it would be possible to compare the survival probability and Hazards Ratio for HF specific hospitalisations or HF specific emergency admissions. This information could bring important insights into the associations between the clusters and the outcomes. It would also be of interest to consider the outcome mortality in this analysis.

Another idea worth exploring is related to studying the temporal evolution of the clusters and analysing the progression and evolution of the disease state over time. Taking into account the work by Vetrano et al. (2020) for a cohort of patients with multimorbidity, the same approach could be applied for a cohort of HF patients. This would allow to tracing the evolution of the clusters and clinical trajectories, which can in turn aid healthcare professionals in defining more personalised treatment and preventive strategies.

Future work should focus on validation in an external cohort, which can be done by replicating this approach in a dataset from a different hospital or healthcare facility. This will be key to determining whether the obtained HF clusters are replicable in other populations and EHR systems.

Finally, considering that the approach used to identify the patient subgroups was unbiased, that is, data-driven with no *a priori* assumptions, it would be interesting to study the adaptability of this approach to other complex conditions where finding condition subtypes could be of use. Some examples include Dementia, COPD and ICM.

Bibliography

R. Navickas, V.-K. Petric, A. Feigl, and M. Seychell. Multimorbidity: What Do We Know? What Should We Do? *Journal of Comorbidity*, 6:4–11, 02 2016. doi: 10.15256/joc.2016.6.72.

M. Rijken, V. Struckmann, M. Dyakova, M. G. Melchiorre, S. Rissanen, and E. van Ginneken. ICARE4EU: Improving Care For People With Multiple Chronic Conditions in Europe. *Eurohealth*, 19(3):29 – 31, 2013.

C. Bähler, C. Huber, B. Brüngger, and O. Reich. Multimorbidity, Health Care Utilization and Costs in an Elderly Community-dwelling Population: A Claims Data Based Observational Study. *BMC health services research*, 15:23, 01 2015. doi: 10.1186/s12913-015-0698-2.

F. Colombo, M. Goñi, and C. Christoph. Addressing Multimorbidity to Improve Healthcare and Economic Sustainability. *Journal of Comorbidity*, 6:21–27, 02 2016. doi: 10.15256/joc.2016.6.74.

N. R. Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. 01 2012. doi: 10.17226/13284.

C. Shivade, P. Raghavan, E. Fosler-Lussier, P. Embi, N. Elhadad, S. Johnson, and A. Lai. A Review of Approaches to Identifying Patient Phenotype Cohorts Using Electronic Health Records. *Journal of the American Medical Informatics Association : JAMIA*, 21, 11 2014. doi: 10.1136/amiajnl-2013-001935.

C. Raherison, E. Ouaalaya, A. Bernady, J. Casteigt, C. Nocent-Eijnani, L. Falque, F. Guillou, L. Nguyen, A. Ozier, and M. Molimard. Comorbidities and COPD severity in a clinic-based cohort. *BMC Pulmonary Medicine*, 18, 07 2018. doi: 10.1186/s12890-018-0684-7.

M. A. Thenganatt and J. Jankovic. Parkinson Disease Subtypes. *JAMA Neurology*, 71(4):499–504, 04 2014. ISSN 2168-6149. doi: 10.1001/jamaneurol.2013.6233.

G. S. Francis, R. Cogswell, and T. Thenappan. The Heterogeneity of Heart Failure. *Journal of the American College of Cardiology*, 64(17):1775–1776, 2014. doi: 10.1016/j.jacc.2014.07.978.

A. Groenewegen, F. Rutten, A. Mosterd, and A. Hoes. Epidemiology of heart failure. *European Journal of Heart Failure*, 22, 06 2020. doi: 10.1002/ejhf.1858.

T. Nagamine, B. Gillette, A. Pakhomov, J. Kahoun, H. Mayer, R. Burghaus, J. Lippert, and M. Saxena. Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. *Scientific Reports*, 10(1), 2020. doi: 10.1038/s41598-020-77286-6.

WHO. Multimorbidity: Technical Series on Safer Primary Care. *Licence: CC BY-NC-SA 3.0 IGO.*, pages 18–28, 2016.

P. Broeiro-Gonçalves. Multimorbilidade e Comorbilidade: Duas Perspectivas da Mesma Realidade. *Rev Port Med Geral Fam*, 31, 06 2015. doi: 10.32385/rpmgf.v31i3.11520.

The Academy of Medical Sciences. Multimorbidity: a Priority for Global Health research. 06 2018.

G. Savarese and L. Lund. Global Public Health Burden of Heart Failure. *Cardiac Failure Review*, 03:7, 04 2017. doi: 10.15420/cfr.2016:25:2.

C. W. Yancy, M. Jessup, B. Bozkurt, J. Butler, D. E. Casey, M. H. Drazner, G. C. Fonarow, S. A. Geraci, T. Horwich, J. L. Januzzi, M. R. Johnson, E. K. Kasper, W. C. Levy, F. A. Masoudi, P. E. McBride, J. J. McMurray, J. E. Mitchell, P. N. Peterson, B. Riegel, F. Sam, L. W. Stevenson, W. W. Tang, E. J. Tsai, and B. L. Wilkoff. 2013 ACCF/AHA Guideline for the Management of Heart Failure. *Circulation*, 128 (16):e240–e327, 2013. doi: 10.1161/CIR.0b013e31829e8776.

B. Bozkurt, A. Coats, and H. Tsutsui. Universal Definition and Classification of Heart Failure. *Journal of cardiac failure*, 27, 02 2021. doi: 10.1016/j.cardfail.2021.01.022.

C. D. Kemp and J. V. Conte. The pathophysiology of Heart Failure. *Cardiovascular Pathology*, 21(5): 365–371, 2012. doi: https://doi.org/10.1016/j.carpath.2011.11.007.

A. Inamdar and A. Inamdar. Heart Failure: Diagnosis, Management and Utilization. *Journal of Clinical Medicine*, 5:62, 06 2016. doi: 10.3390/jcm5070062.

S. Anker, J. Butler, G. Filippatos, M. Khan, N. Marx, C. Mbbs, S. Schnaidt, A. Ofstad, M. Brueckmann, W. Jamal, E. Bocchi, P. Ponikowski, S. Perrone, J. Januzzi, S. Verma, and M. Böhm. Effect of Empagliflozin on Cardiovascular and Renal Outcomes in Patients With Heart Failure by Baseline Diabetes Status: Results From the EMPEROR-Reduced Trial. *Circulation*, 143, 11 2020. doi: 10.1161/CIRCULATIONAHA.120.051824.

J. Banda, M. Seneviratne, T. Hernandez-Boussard, and N. Shah. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annual Review of Biomedical Data Science*, 1, 07 2018. doi: 10.1146/annurev-biodatasci-080917-013315.

N. 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 : JAMIA*, 20, 06 2012. doi: 10.1136/amiajnl-2011-000681.

M. Tayefi, P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen. Challenges and Opportunities Beyond Structured Data in Analysis of Electronic Health Records. *WIREs Computational Statistics*, 2021. doi: 10.1002/wics.1549.

WHO. International Statistical Classification of Diseases and Related Health Problems (ICD) , 2018. URL `https://www.who.int/standards/classifications/classification-of-diseases`.

S. Esteban, M. Tablado, R. Ricci, S. Terrasa, and K. Kopitowski. A Rule-Based Electronic Phenotyping Algorithm for Detecting Clinically Relevant Cardiovascular Disease Cases. *BMC Research Notes*, 10, 07 2017. doi: 10.1186/s13104-017-2600-2.

W.-Q. Wei, P. Teixeira, H. Mo, R. Cronin, J. Warner, and J. Denny. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association : JAMIA*, 23, 09 2015. doi: 10.1093/jamia/ocv130.

H. Alzoubi, R. Alzubi, N. Ramzan, D. West, T. Al-Hadhrami, and M. Alazab. A Review of Automatic Phenotyping Approaches using Electronic Health Records. *Electronics*, 8:1235, 10 2019. doi: 10.3390/electronics8111235.

J. Pathak, A. N. Kho, and J. C. Denny. Electronic Health Records-Driven Phenotyping: Challenges, Recent Advances, and Perspectives. *Journal of the American Medical Informatics Association*, 20 (e2):e206–e211, 12 2013. doi: 10.1136/amiajnl-2013-002428.

K. Kreimeyer, M. Foster, A. Pandey, N. Arya, G. Halford, S. Jones, R. Forshee, M. Walderhaug, and T. Botsis. Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: a systematic review. *Journal of Biomedical Informatics*, 73, 07 2017. doi: 10.1016/j.jbi.2017.07.012.

R. Carroll, A. Eyler, and J. Denny. Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis. *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011: 189–96, 01 2011.

J. Pestian, K. Cohen, B. Glass, H. Greiner, K. Holland, S. Standridge, R. Arya, R. Faist, D. Morita, F. Mangano, B. Connolly, and T. Glauser. Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning. *Biomedical Informatics Insights*, 8:11, 05 2016. doi: 10.4137/BII.S38308.

J. Ho, J. Ghosh, S. Steinhubl, W. Stewart, J. Denny, B. Malin, and J. Sun. Limestone: High-throughput Candidate Phenotype Generation via Tensor Factorization. *Journal of Biomedical Informatics*, 52, 12 2014. doi: 10.1016/j.jbi.2014.07.001.

E. Bleecker, W. Moore, W. Busse, M. Castro, K. F. Chung, W. Calhoun, S. Erzurum, B. Gaston, E. Israel, D. Curran-Everett, and S. Wenzel. Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data. *The Journal of allergy and clinical immunology*, 133, 02 2014. doi: 10.1016/j.jaci.2013.11.042.

A. Gordon. *Classification*. 1999.

A. Jain. Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31:651–666, 06 2010. doi: 10.1016/j.patrec.2009.09.011.

H. Kashima, J. Hu, B. Ray, and M. Singh. K-means clustering of proportional data using L1 distance. pages 1 – 4, 01 2009. doi: 10.1109/ICPR.2008.4760982.

S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. pages 787–788, 01 2007. doi: 10.1145/1277741.1277909.

B. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster Analysis*, volume 5th. 01 2011. ISBN 9780470978443. doi: 10.1002/9780470977811.

O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46:243–256, 01 2013. doi: 10.1016/j.patcog.2012.07.021.

P. Rousseeuw. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987. doi: 10.1016/0377-0427(87)90125-7.

T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1): 1–27, 1974. doi: 10.1080/03610927408827101.

D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

D. Vetrano, A. Roso-Llorach, S. Fernández-Bertolín, M. Clavero, C. Violan, G. Onder, L. Fratiglioni, A. Calderón-Larrañaga, and A. Marengoni. Twelve-year clinical trajectories of multimorbidity in a population of older adults. *Nature Communications*, 11:3223, 06 2020. doi: 10.1038/s41467-020-16780-x.

C. Violan, A. Roso-Llorach, Q. Foguet-Boreu, M. Clavero, M. Pons-Vigués, E. Pujol-Ribera, and J. Valderas. Multimorbidity patterns with K-means nonhierarchical cluster analysis. *BMC Family Practice*, 19, 07 2018. doi: 10.1186/s12875-018-0790-x.

M. Vandromme, T. Jun, P. Perumalswami, A. Branch, and I. Ii. Automated phenotyping of patients with non-alcoholic fatty liver disease reveals clinically relevant disease subtypes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25:91–102, 01 2020.

T. Ahmad, M. Pencina, P. Schulte, E. O'Brien, D. Whellan, I. Piña, D. Kitzman, K. Lee, C. O'Connor, and G. Felker. Clinical Implications of Chronic Heart Failure Phenotypes Defined by Cluster Analysis. *Journal of the American College of Cardiology*, 64:1765–1774, 10 2014. doi: 10.1016\/j.jacc.2014.07.979.

C. Gulea, R. Zakeri, and J. Quint. Model-based comorbidity clusters in patients with heart failure: association with clinical outcomes and healthcare utilization. *BMC Medicine*, 19, 01 2021. doi: 10.1186/s12916-020-01881-7.

S. Shah, D. Katz, S. Selvaraj, M. Burke, C. Yancy, M. Gheorghiade, R. Bonow, C.-C. Huang, and R. Deo. Phenomapping for Novel Classification of Heart Failure With Preserved Ejection Fraction. *Circulation*, 131, 11 2014. doi: 10.1161/CIRCULATIONAHA.114.010637.

T. Ahmad, L. Lund, P. Rao, R. Ghosh, P. Warier, B. Vaccaro, U. Dahlstrom, C. O'Connor, G. Felker, and N. Desai. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. *Journal of the American Heart Association*, 7:e008081, 04 2018. doi: 10.1161/JAHA.117.008081.

A. Ahmad and S. Khan. A survey of State-of-the-Art Mixed Data Clustering Algorithms. *Ieee Access*, 7: 31883–31902, 03 2019. doi: 10.13140/RG.2.2.17863.55209.

A. Foss, M. Markatou, and B. Ray. Distance Metrics and Clustering Methods for Mixed-Type Data, 04 2019.

S. Singh, S. Karkare, S. Baswan, and V. Singh. Agglomerative Hierarchical Clustering Analysis of co/multi-morbidities. 07 2018.

J. Gower. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27:857–871, 12 1971. doi: 10.2307/2528823.

G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, 8:54776–54788, 2020. doi: 10.1109/ACCESS.2020.2980942.

I. Jolliffe. *Principal Component Analysis*, pages 1094–1096. 01 2011. ISBN 978-3-642-04897-5. doi: 10.1007/978-3-642-04898-2_455.

H. Abdi and D. Valentin. Multiple Correspondence Analysis. *Encyclopedia of Measurement and Statistics*, 01 2007.

F. Husson, J. Josse, and S. Lê. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25, 03 2008. doi: 10.18637/jss.v025.i01.

J. Verdonschot, M. Merlo, F. Dominguez, P. Wang, M. Henkens, M. Adriaens, M. Hazebroek, M. Masè, L. Escobar, R. Cobas Paz, K. Derks, A. Wijngaard, I. Krapels, H. Brunner, G. Sinagra, P. Garcia-Pavia, and S. Heymans. Phenotypic clustering of dilated cardiomyopathy patients highlights important pathophysiological differences. *European Heart Journal*, 42, 11 2020. doi: 10.1093/eurheartj/ehaa841.

C. Cramer, M. Porter, H. Sayama, L. Sheetz, and S. Uzzo. *Network Literacy: Essential Concepts and Core Ideas*. 03 2015.

C. Hidalgo, N. Blumm, A.-L. Barabasi, and N. Christakis. A Dynamic Network Approach for the Study of Human. *PLoS computational biology*, 5:e1000353, 05 2009. doi: 10.1371/journal.pcbi.1000353.

V. Fionda and L. Palopoli. Biological network querying techniques: Analysis and comparison. *Journal of computational biology : a journal of computational molecular cell biology*, 18:595–625, 03 2011. doi: 10.1089/cmb.2009.0144.

T. Clark, M. Bradburn, S. Love, and D. Altman. Survival Analysis Part I: Basic Concepts and First Analyses. *British Journal of Cancer*, 89:232–8, 08 2003.

V. Bewick, L. Cheek, and J. Ball. Statistics review 12: Survival analysis. *Critical care (London, England)*, 8:389–94, 11 2004. doi: 10.1186/cc2955.

A. Waljee, A. Mukherjee, A. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, and P. Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3, 08 2013. doi: 10.1136/bmjopen-2013-002847.

M. Azur, E. Stuart, C. Frangakis, and P. Leaf. Multiple Imputation by Chained Equations: What is it and how does it work? *International journal of methods in psychiatric research*, 20:40–9, 03 2011. doi: 10.1002/mpr.329.

G. Yilmaz and H. Shaikh. Normochromic Normocytic Anemia. *StatPearls [Internet]*, 12 2020.

OECD/WHO. *Overweight and obesity*. 2020. doi: /https://doi.org/10.1787/a47d0cd2-en.

European Society of Cardiology. Acute and Chronic Heart Failure Guidelines. ESC Clinical Practice Guidelines. 2016.

M. Szumilas. Explaining Odds Ratio. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Académie canadienne de psychiatrie de l'enfant et de l'adolescent*, 19:227–9, 08 2010.

V. Rudrapatna and A. Butte. Opportunities and challenges in using real-world data for health care. *Journal of Clinical Investigation*, 130:565–574, 02 2020. doi: 10.1172/JCI129197.