Using Markov Chains and Temporal Alignment to Identify Clinical Patterns in Dementia

Maria Luísa Marote e Costa

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Susana de Almeida Mendes Vinga Martins
Prof. Andreia Sofia Monteiro Teixeira

Examination Committee

Chairperson: Prof. Ana Luísa Nobre Fred
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Prof. Pedro Tiago Gonçalves Monteiro

December 2021

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.

Preface

The work presented in this dissertation was performed at Hospital da Luz Lisboa (Lisbon, Portugal), during the period March-October 2021, under the supervision of Prof. Andreia Sofia Monteiro Teixeira, within the scope of the IntelligentCare project LISBOA-01-0247-FEDER-045948 that is co-finnanced 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. Susana Vinga.

Acknowledgments

I would like to thank everyone who supported me during the adventure that was completing my Masters, which comes to an end with the conclusion of this dissertation.

Firstly, I would like to express my deep gratitude to Prof. Sofia Teixeira for all her guidance, support and fellowship during this long process. All your incentives, advice and hours invested during the development of this work were essential for its success and it would not be possible without you. To Prof. Susana Vinga and Prof. Alexandra Carvalho, I want to thank for all the support and feedback throughout the progression of this project.

I would also like to thank Dr. João Colaço from Hospital da Luz who provided huge medical expertise during the whole process and always managed to make time to help. To Dr. Raquel Gouveia, thank you for your help and suggestions. To Miguel Froes, thank you for your cooperation and advice.

To my mother, father, brother, grandparents and stepfather, thank you for your endless support and for never letting me doubt myself, always encouraging me to do better. A special deep gratitude goes out to my mother, who always guided me through my setbacks and helped me back up. To my aunt and uncle, who treated me as a daughter through these five years of university, I am forever grateful.

To my friends and colleagues that I met during these last five years, this journey would not have been the same without you. A special thank you to my ”Marias”, Maria Jacinto and Maria Teresa Marcelino, for your friendship, laughter and fellowship. And last but not least, to my friends of a lifetime, I want to express my deepest gratitude for your friendship and existence in my life. We grew and learned so much from one another, even if not always close, I am sure we will always be a constant in each others paths. To my number ones, ”Ini” and ”Alinho”, a special thank you for everything.

Abstract

In the last decades, big data and advanced analytics have enabled public and private sectors to optimize their performance through personalized targeting and traits. When it comes to the healthcare sector, this becomes even more important as the complexity of a patient, in terms of their comorbidities, increases as well as the need for a more integrated and patient-centered treatment plan. In this work, we focus on understanding key phenotypes and clinical pathways of patients with multimorbidity suffering from Dementia, a disease that can result from very heterogeneous factors and has the potential of becoming more prevalent as the population ages. In this dissertation we present a set of methods which allow us to identify phenotype patterns and to find recurrent patterns of medical consults within the entire cohort, as well as to stratify patients into subgroups that exhibit similar patterns of interaction. With Markov Chains we are able to identify the most prevailing medical consults attended by Dementia patients, as well as recurring transitions between different medical speciality consults. With AliClu, the algorithm used to stratify patients, we successfully identify patient subgroups which present similar medical consult activity and also identify similar patterns of interaction within these subgroups. A phenotype analysis per cluster obtained allows to identify distinct patterns and characteristics. This pipeline provides a tool to identify prevailing clinical pathways of medical consultations within the dataset, as well as the most common transitions between medical specialities within Dementia patients. This methodology, alongside demographic and phenotypic data, has the potential to provide early signalling of the most likely clinical pathways and serve as a support tool for health providers on deciding the best course of treatment, considering a patient centered approach.

Keywords

Multimorbidity; Dementia; Markov Chains; Temporal Sequence Alignment; Clustering.

Resumo

Ao longo das últimas décadas, *big data* e métodos analíticos avançados têm permitido aos setores públicos e privados otimizar o seu desempenho através de soluções personalizadas. No que toca ao setor da saúde, estas soluções tornam-se ainda mais importantes à medida que a complexidade dos pacientes, em termos de comorbidades, aumenta. Por sua vez, isto leva à necessidade de se criar um plano de tratamento mais integrado e centrado no paciente. Neste estudo, o foco está em detetar padrões chave de fenótipos e atividade clínica de pacientes com multimorbilidade que sofrem de Demência. Esta é uma doença que pode advir de diversos fatores com potencial para se tornar cada vez mais incidente à medida que a população envelhece. Nesta dissertação, apresentamos um conjunto de métodos que permitem identificar fenótipos chave e padrões de interacções com consultas de especialidade, assim como estratificar a população de acordo com padrões semelhantes de interação com o hospital. Recorrendo a cadeias de Markov conseguimos identificar os tipos de consultas mais prevalentes, assim como as transições mais recorrentes entre especialidades. Com recurso ao algoritmo usado para a estratificação, nomeadamente o AliClu, identificámos com sucesso grupos de pacientes que apresentam padrões de atividade semelhantes, o que nos permitiu distinguir padrões de interação semelhantes dentro de cada agrupamento. Uma análise de fenótipos por cada conjunto obtido permitiu o reconhecimento de padrões que os diferenciam. Esta metodologia, em conjunto com dados demográficos e fenotípicos, tem o potencial de fornecer uma sinalização precoce da atividade clínica mais provável, servindo como ferramenta de apoio aos profissionais de saúde para decidir qual o melhor plano de tratamento, considerando uma abordagem centrada no paciente.

Palavras Chave

Multimorbilidade; Demência; Cadeias de Markov; Alinhamento de sequências temporais; Estratificação.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

AFib	Atrial Fibrillation
BPH	Benign Prostatic Hyperplasia
CM	Cardiomyopathy
COPD	Chronic Obstructive Pulmonary Disease
CKD	Chronic Kidney Disease
DLB	Dementia with Lewy bodies
DTW	Dynamic Time Warping
EHR	Electronic Health Records
EMR	Electronic Medical Records
g	Gap penalty
GFM	General and Family Medicine
HMM	Hidden Markov Models
HLL	Hospital da Luz Lisboa
HTN	Hypertension
IT	Information Technology
ML	Machine Learning
MM	Multimorbidity
NLP	Natural Language Processing
NW	Needleman-Wunsch

k	Number of clusters
Ob-Gyn	Obstetrics and Gynaecology
OR	Odds Ratio
ONC	Office of the Natural Coordinator
OECD	Organisation for Economic Co-operation and Development
ORL	Otorhinolaryngology
PE	Prefix-encoded
SS	Silhouette Score
TNW	Temporal Needleman-Wunsch
Tp	Temporal penalty
TM	Transition Matrix
T2DM	Type 2 diabetes
VD	Vascular Dementia
WHO	World Health Organization

1

Introduction

Contents

1.1 Motivation

In the last decades, big data and advanced analytics have enabled public and private sectors to optimize their performance through personalized targeting and traits. When it comes to healthcare, this becomes even more important as the complexity of a patient in terms of their comorbidities increases, as well as the need for a more integrated treatment plan.

The healthcare industry is one of the sectors that can most benefit from big data analysis. According to Dash et. al [1], big data in healthcare comprises hospital records, patient medical records and results of medical examinations, as well as biomedical research, among other sources of information about the patients. To aim for personalized medicine, it is necessary to manage and analyse these data strategically. This is crucial when addressing complex patients with multiple comorbidities, even more when these include chronic diseases. Multimorbidity can be defined as more than one chronic or long-term condition. Despite its prevalence in the population, specially elder population, there is still a long way to understanding its patterns and the best way to plan treatment. Guidelines for care providing are still very much focused on single-disease patient models. It is imperative to switch focus onto a more patient-centred model addressing all patient's needs, offering an integrated and more coordinated care. As mentioned in other studies [2], the European Commission has worked on promoting innovation and research to improve patient-centred integrated care, targeting patients with multimorbidity. Thus, analysing complex heterogeneous groups of patients and finding significant patterns amongst the population can represent an important first step in this direction.

As the care providing system improves, the population in general will tend to live longer. With this growth in ageing, the incidence of multimorbidity will tend to increase, leading to an overload of the healthcare system. Given the fact that these patients need regular attention, it raises a need to prepare treatment plans to serve their needs. In order to identify certain characteristics of this fraction of the population, a crucial approach is to resort on Electronic Medical Records (EMR). The rise in the availability of these EMR allows us to store all types of information – structured and unstructured – which can be used to gather insightful awareness for the future of the healthcare sector. The growing attention towards the process of data mining supports this search for more integrated information. Data mining is focused on discovering patterns and correlations within heterogeneous large data sets, resorting to a broad range of techniques which can include machine learning and artificial intelligence.

Despite the advantages of exploring the available healthcare data and using it to improve present outcomes, there are also challenges associated to this process. The amount of information that exists, associated to the fact that it can be very heterogeneous, encountering both structured and unstructured data, can pose a challenge for the data mining process. For instance, in the healthcare sector, the way the information is gathered may vary depending on several factors. Some of these factors include the fact that some fields may be wrongly introduced by health professionals or not introduced at all, which induces possible noise in the data. This leads to a need of a careful preprocessing phase to guarantee the data being explored is well formatted.

Additionally, patients suffering from multimorbidity can be divided in very heterogeneous groups, which flags the importance of performing a phenotype screening when exploring their data. Characterizing a group of patients according to several phenotypes, such as age, gender, chronic diseases, within others, allows the identification of specific attributes and patterns. This is a great contribution to build tools to support treatment planning. As an example, phenotype analysis coupled with a patient stratification technique can provide a great support when deciding what is the best treatment approach, depending on which subset the patient is inserted.

1.2 Objectives and contributions

The main goal of this thesis is to provide a pipeline that allows us to analyse and to identify patterns within a complex cohort of patients suffering from multimorbidity in which Dementia is included. In this document we show a set of methods which allows us to find recurrent patterns of medical consults within the entire cohort, as well as to stratify patients into subgroups that exhibit similar patterns of interactions with the hospital.

The work consisted in two main phases. First, in order to find certain characteristics of these patients, we performed a phenotype screening of the population in hand, including a study of the various chronic diseases present in the population.

The next phase was centered in the activity of the patients considering their various medical consults throughout the considered time window. This clinical pathway analysis was tackled resorting to two distinct methods. In order to identify recurring patterns within the population, we started by focusing on the transitions between medical specialities and the probability of each one of them. Finally, to determine the existence of certain sub-groups within the chosen cohort based on consult attendance, a clustering approach was implemented, complementing this with a phenotype screening within each subset.

The information on the most recurrent characteristics and patterns of clinical pathways relative to Dementia patients, alongside demographic and phenotypic data, has the potential to provide early signalling of the most likely clinical pathways. Stratifying the patients based on their activity allows us to detect different subgroups of patients with similar characteristics, promoting an easier determination of the best course of treatment. Thus, the adaptation of the pipeline to other cohorts may serve as a support decision systems tool for medical practitioners to provide a more patient-centered care, considering a patient as a whole and not focusing only on a certain problem.

An article with early results of the work presented in this dissertation, entitled ”Using Markov Chains and Temporal Alignment to Identify Clinical Patterns in Dementia” was presented in the 12th edition of a Portuguese informatics conference called INForum. A second article is now under development to be submitted to a journal.

1.3 Thesis Outline

This Master's Dissertation is organized as follows:

- Chapter 2: This introductory chapter serves the purpose of presenting all concepts related to the work developed. First of all, base concepts such as Electronic Health Records and Multimorbidity are introduced, followed by the presentation of the study cohort used, finalizing with related work.
- Chapter 3: This division details the methods used to fulfill the outlined goals of this project, starting by presenting the data in more detail and clarifying what was explored with the initial phenotype screening process. Next, the methods used for the clinical pathway analysis, specifically Markov Chains and the AliClu clustering algorithm, are presented in detail.
- Chapter 4: This chapter shows the results obtained for all the stages of this project. It starts with the visualization of the data considering the different analysed phenotypes. Moving on to the clinical pathway analysis sector, the results of the Markov Chains approach are presented, followed by the results of the clustering algorithm used, which is divided in four parts. The pre processing analysis is shown, followed by the parameter optimization process, then the final clusters obtained are presented followed by the corresponding phenotype screening per cluster. In addition, a discussion of the obtained results as well as a parallelism to what exists in the literature is done.
- Chapter 5: The final chapter summarizes the main outcomes of the performed work and refers to how the methods used can have a big influence in moving a step further in how healthcare data can be used to improve the industry and the care providing system. Finally, improvements that can still be done and future work in this context are discussed.

2

Main concepts and related work

Contents

2.1 Electronic Health Records

The rise of Information Technology (IT) in healthcare brought along the introduction of Electronic Medical Records. The Office of the Natural Coordinator (ONC) for Health IT [3] defines EMRs as a digital version of paper charts in a health professional's office, containing the medical history, as well as the treatment history of patients in one health facility. However, it is important to point out the definition of Electronic Health Records (EHR) as well. EMRs and EHRs are many times used as synonyms, nonetheless, there is a difference. While it is difficult for an EMR to transcend its practice, causing a non-availability of a patient's information outside that specific organization, an EHR is made to go beyond the healthcare facility that is responsible for it. Thus, an EHR can reach out to a broader range of experts, or even laboratories, allowing any clinician that comes across a certain patient to access records of the previous treatments and specialists involved in that patient's care. As stated by the ONC for Health IT, EHRs focus on the patient's total health, moving beyond standard clinical data collective in the healthcare provider's office, transcending to a broader view on a patient's care [3].

When considering the analysis of these electronic records, it is important to understand their structure. When clinical data on patients is generated during the diagnosis process it can be stored in structured or unstructured formats within the EHR system. The former can be coded medical data, such as diagnostic or procedural codes, laboratory exam results or vital signs, while the latter is everything that is stored as text or images as doctor's notes on a patient's progress, discharge and admission notes, as well as more complex exam results. Additionally, there is compelling variability between how different health institutions and even providers input patient data into the EHR, not to mention the challenge regarding the accuracy of the information that is stored, which can vary considerably, as stated by previous studies [4]. Thus, the treatment of structured and unstructured EHR may differ in the methods used.

2.2 Electronic Phenotyping

EHRs and EMRs can be associated to the appropriate evidence-based tools, having potential when it comes to developing clinical decision support systems for healthcare providers, which could have a huge impact on patient diagnosis, as well as in quality measures. Furthermore, these can be used to assist in a broad range of experiments regarding patient information. For instance, discovering phenotype-genotype associations, establishing clinical trial protocols, automating drug event detection, as well as prevention, accelerate the research on precision medicine [5], and electronic phenotyping. The latter can be defined as finding patients with specific conditions or outcomes [4], representing one of the major approaches concerning the utilization of EHR.

The process of phenotyping can be used in a broad range of applications when it comes to exploring medical data, for instance screening the most adherent medications within a certain set of patients, or identifying prevailing chronic conditions within a population, exploring patient reaction to a certain line of treatment, within others. Several studies have been conducted with several purposes resorting to electronic phenotyping techniques. Banda et al. [4] mentions that phenotyping can be used in (i) cross-sectional studies, such as epidemiological studies and tracking adherence to treatment guidelines epidemiological studies, (ii) cohort-control and case-control analysis, like identifying certain risk factors within a subset of patients, (iii) genome- and phenotype-wide association studies, linking genomic data with phenotypes gathered from EHR, and (iv) experimental studies, determining eligibility to participate in clinical trials.

As a result of the complexity and heterogeneity of EHR, electronic phenotyping can involve quite some challenges. Due to the presence of both structured and unstructured data within EHR, coupled to the variability of data storage across different health institutions, the process of electronic phenotyping becomes complex, enabling the existence of a standardized tool for this purpose. Instead, strong methods, that allow to deal with the heterogeneity between patient records and data types, leading to the extraction of significant phenotypes, are developed according to the specific situation in hands. The creation of phenotyping methods that can be used on several grounds represents a challenge.

Banda et al. [4] identifies several phenotyping methods, such as rule-based, Natural Language Processing (NLP), Machine Learning (ML) and hybrid frameworks. The first one is indicated when the aim is to deal with phenotypes that have clear medical codes associated, such as diagnosis ones, and involves clinicians stating inclusion and exclusion criteria based on several data elements, like medication, diagnosis codes, medical procedures and lab results. NLP is specially useful when dealing with unstructured data of EHR, such as text from clinical notes or radiology reports, however represents many challenges due to the amount of information registered by text, making it harder to reliably extract phenotypes. Furthermore, Banda et al. [4] mentions, however, that NLP techniques have been growing and now represent a great part of electronic phenotyping and is many times used alongside other methods like rule-based or ML. Regarding ML techniques, which can vary depending on the quality of the data. For instance, standard ML methods like Naive Bayes and Support Vector Machines all need labeled data, which requires it to be well homogenized and stored, which is not always the case. However, the use of noisy imperfect data have also proved to allow extraction of significant phenotypes [6] [7]. Finally, it is also possible to resort to unsupervised ML methods, this is, extracting phenotypes from data without any labels. For instance, Ho et al. [8] used a tensor factorization method to generate phenotypes in the form of clusters, where each one represented a certain medical condition, resorting to medical specialists to validate the results.

2.3 Multimorbidity

A broad range of definitions for multimorbidity (MM) can be found in literature, but according to Almirall et al. [9], there are two definitions which are mostly used. These are that MM can be defined as “more than one or multiple chronic or long-term diseases/conditions” or as “more than one or multiple diseases or conditions”, the latter being distinct from the former since it does not specify the nature of the disease (if long-term or chronic).

Ageing is the most persistent factor mentioned in the literature when it comes to the increase in MM risk, having several studies mentioned a direct association between age and prevalence of MM. It has also been pointed out, through systematic reviews on studies concerning MM, that most individuals older than 65 years live with MM [10]. However, it is important to advert to the fact that, despite of the increase of this prevalence along with ageing, it is not a condition only affecting the elder population [11]. The increase of MM amongst the population has also been associated to lifestyle changes, health seeking behavior and the environment [12].

Furthermore, a study carried out in Finland [13] came to the conclusion that the incidence on MM is highly influenced by several clinical and lifestyle risk factors, both on patients that suffer from a chronic disease, as well as on those who do not. These risk factors included smoking, high body mass index and physical inactivity, for both men and women, while high blood pressure and low education were considered additional factors for men.

Multimorbidity can very much affect several entities, including the patient itself, as well as care providing facilities (i.e. hospitals). It can impact an individual’s day-to-day life, his physical resistance as well as their quality of life, in a way that may lead to depressive states, which will increase even more their medication intake [11]. MM patients are ones with complex care needs, being very costly and challenging to manage, in a way that they are more frequently admitted in hospitals or clinics, registering longer hospital stays, as well as the fact that the range of specialists to attend is broader than usual.

Despite the prevalence of MM in the population, there is still a long way to understanding its patterns and the best way to face treatment. Guidelines for care providing are still very much focused on single-diseased patients, which makes it imperative to use nowadays technology to change this focus onto a more patient-centred model, that is prepared to address and consider all of the patient’s needs and pathologies, providing an integrated and more coordinated care. As stated by Rijken et al. [14], the European Commission has worked on promoting innovation and research to improve patient-centred integrated care, targeting patients with MM. However, there are still some topics surrounding MM, such as epidemiology, risk factors, how to better prevent and manage it and how to optimize care delivery, that should be explored in order to understand it the best way possible.

Research in this area has received more and more attention throughout the years, however, it has been mostly focused on the prediction, prevention and management of disorders independently from one another [12]. The Academy of Medical Sciences published an international policy report, in 2018, that aims at evaluating the growing MM issue as a global health challenge, summarizing the available evidence on MM, as well as highlighting key gaps [15]. Throughout this report, it is recommended a shift in goals when studying MM, in order to better understand trends and patterns of MM across the world, the burden related to common clusters of conditions, its determinants, how best to prevent the development of MM, how to maximize benefits, limit risks and organise healthcare systems in order to better manage the treatment of patients suffering from MM.

2.3.1 Multimorbidity analysis

The identification of multimorbidity patterns is rising as a critical step in the development of healthcare services that are sensitive to a patient's health needs. In order to try to better understand MM, its causes and consequences, its patterns, prevalence in certain age groups, as well as the existing relationships between co-existing diseases, among others, several methods have been implemented and tested along the years.

Ng et al. [16] published a systematic review of analytical methods used to identify patterns of MM. The authors performed a review on the analytical methods in epidemiological studies on MM, identifying five distinct methods for understanding the nature and patterns of MM. They found heterogeneity within each method, mainly due to the proximity measures used to form clusters, which differed from study to study. The focus was on methods quantifying non-random MM and assessing differences in MM patterns within the general population, in opposition to only focusing on a group of people with a certain condition. The findings that arise from these types of studies can be of great importance, since they may generate new hypothesis on possible shared biologic processes and facilitate quantification of the impact that multimorbidity shows regarding quality of life and health-related outcomes. This, in turn, can be useful for future studies focusing on improving prevention, treatment and care of patient suffering from multiple conditions. The methods identified to assess MM patterns by Ng et al. [16] included factor-analysis method, hierarchical-clustering methods, unified-clustering algorithm, multiple correspondence analysis and finally, network and cluster analysis.

Furthermore, Hassaine et al. [12] mentions two main types of studies used in the past to investigate MM, which they refer to as ”pairwise methods” and ”factorization methods”. The former refers to methods based on neural networks, where nodes are viewed as diseases and edges represent the extent to which each disease pair is connected. These neural networks can then be used to characterize MM pathways, however, these methods may be biased, due to their inability to consider conditional independence between diseases. Hence, the need for alternative methods that consider the whole dynamics among co-existing diseases, not just pairs of them. The basis of factorisation methods is to decompose a matrix, which denotes the metrics between patients and diseases, into a number of MM patterns, each consisting of a disease cluster and its expression amongst patients. Initially, these methods did not consider the temporal dimension of the patterns, however, methods which do consider have also been studied.

Following the same line of thought of moving past the analysis of simply considering pairs of diseases, Hassaine et al. [17] refers to probabilistic methods as another approach. For instance, latent class growth modelling is mentioned as an example for identifying clusters of MM trajectories. Hidden Markov Models (HMM) are also mentioned as an approach to modelling these trajectories, since these can incorporate time while learning the progression of a patient's health. Hassaine et al. [17] also points out factorisation methods, describing factorisation as an assumption that each patient's health record is a combination of several factors, present in the whole population, but with some variability from individual to individual, which is a consequence of the extent to which these factors are expressed in each one. With this information, it is possible to construct a matrix denoting the metrics between patients and diseases (factors) to start off the model, as mentioned above.

Moreover, Zhou et al. [18] developed a temporal phenotyping approach, in order to account for the temporal dimension of MM, constructing a longitudinal patient matrix for each patient consisting of both a feature and time dimension, resourcing to the available electronic medical records. What differentiates this approach from other matrix involving approach is that each patient is represented by a matrix, instead of a vector, allowing to handle missing data, since these matrices may have missing entries. It is important to have a method that can work with missing data since many EHR have that problem associated.

2.4 Dementia

Given the rising importance of learning how to deal with multimorbidity patients, treating them as a whole, not merely focusing on dealing with a specific condition alone, the goal was to pick an heterogeneous group of patients to put this study in motion.

According to the World Health Organization (WHO), Dementia is defined as a syndrome of chronic or progressive nature, leading to a degradation in cognitive function and affecting memory, the capacity to process thought, orientation, language, judgement, ability to calculate and learn, within others [19]. Furthermore, it is also stated that this disease is the seventh leading cause of death within other diseases, representing one of the most prevailing causes of dependency and impairment amongst the elder population worldwide [19]. As the population ages, the incidence of this disease will tend to grow. The 2019 report from the Organisation for Economic Co-operation and Development (OECD) states that ageing persists as the most considerable risk factor for Dementia, indicating that the prevalence of this diseases stands at 2.3% among patients between the ages of 65 and 69 years old, while considering an age group higher than 90 years old this prevalence rises to 42% [20]. Portugal specifically, according to this report, stands as the fourth country with higher incidence of this illness, presenting an average of around twenty-one cases per 1 000 population [20], being estimated that by 2050 this incidence will increase to forty cases. The growing prevalence of Dementia, coupled with the fact that a lack of awareness and understanding of this disease, which leads to stigma creation around it and results in barriers to diagnosis and care has been pinpointed [19], represented one of the reasons why this cohort was flagged.

Considering that this study has as one of its main goals to study patients with MM, in order to better understand and manage it, it was important to decide on a group of patients with high tendency of developing comorbidities. It has been evidenced that there is a high incidence of comorbid medical conditions within patients with Dementia [21], being estimated that these can suffer from two to eight additional chronic illnesses [22]. Incidence of certain co-existing diseases in the population with Dementia may even aggravate the patients' conditions. For instance, as stated by Bunn et al. [21], type 2 diabetes may accelerate the cognitive decline of a patient that suffers from Dementia. This high incidence of additional chronic diseases associated to these patients brings many challenges to the healthcare system since under-diagnosis and treatment of Dementia may result from an accelerated advancement towards deteriorated cognitive and functional states due to co-existing illnesses [22] [21]. Additionally, these patients show high utilization of health services, representing a compelling portion of costs associated to healthcare concerning the elderly population, being that the prevalence of comorbidities aggravate this usage, increasing hospital stay, healthcare costs and mortality rates for hospitalized patients [22]. Given all these challenges associated to the incidence of Dementia in the population, together with other chronic conditions, it is important to move towards a better understanding of certain relationships between co-existing illnesses, shifting the focus to treating these patients from a global perspective.

Moreover, Dementia is a disease that can arise from several diseases with distinct etiologies and pathophysiologies [23] [22], being associated to a phenotypic heterogeneity that can lead to a wide range of symptoms, progression rates and disease trajectories, as well as onset ages [24]. Different forms of Dementia include Alzheimer's disease, which is the most common form (approximately 60 to 70% of cases), Vascular Dementia, Dementia with Lewy bodies, frontotemporal Dementia, or it can also result from an event of a stroke, an infectious disease, repetitive physical brain injuries or nutritional deficiencies [19]. It is also important to point out the fact that a disproportionate impact on female patients has been registered regarding this disease [19].

Vascular Dementia (VD) is defined as the deterioration of memory and cognitive functioning resulting from cerebrovascular disease [25]. This type of Dementia is the second leading cause of cognitive impairment, following Alzheimer's disease and can be very heterogeneous in terms of phenotype and pathogenic mechanisms [25]. Cerebrovascular disease may result from other medical conditions, such as cardiovascular risk factors, which may also lead to this deterioration caused by VD. On the other hand, Dementia with Lewy bodies (DLB) is characterized by the accumulation of a protein in Lewy bodies and neurites and resembles with Parkinson's disease, in such a way that DLB can be confused with Dementia developed as a progression from Parkinson's disease itself [26].

Besides these more specific diseases that can easily lead to a Dementia state, there are other risk factors that can be associated with cognitive decline, such as heart disease, depression, obesity and sleep. It has been pointed out in the literature that an association between depression and Dementia exists, specially when considering earlier-life depression, which has been coupled to an increase risk of developing Dementia [27]. Late-life depression has also proven to be linked to Dementia states, however, there are always other factors that may influence [27]. It is also stated that vascular diseases have the strongest evidence connecting depression to Dementia, presupposing that cerebrovascular disease perpetuates some depressive syndromes [27], existing a number of studies reporting that vascular lesions may contribute to depression in late life [28].

Given all these challenges and heterogeneity associated with Dementia, this group of patients was chosen and gathered to put this study in practice, in order to enable pattern identification within these patients, both in terms of phenotypes, as well as in terms of clinical pathway.

2.5 Clinical Pathway Analysis

Discovering clinical pathway patterns represents an interesting approach for acknowledging the structure and dynamics of healthcare guidelines, supporting health providers to better guide treatment and diagnostic pathways. It has been pointed out in the literature throughout the years that the healthcare industry should adopt clearly delineated clinical pathways for patients and that these should be regularly improved and updated [29] [30] [31]. Huang et al. [29] defines a Clinical Pathway as a set of treatment and therapy activities representing the various steps to reach a certain state in a patients care providing path.

Lin et al. [32] identifies four essential factors of a clinical pathway, which are a timeline, the clinical activity type and corresponding interventions, outcome criteria and variance records, which compares the planned outcome with the actual one. Discovery of patterns in clinical pathways can be a challenging task due to the diversity of treatment plans due to the presence of various treatment activities, which in turn may have large variability that depend on time, location and patient characteristics factors.

Several data mining approaches have been purposed to face the challenge of uncovering clinical pathway patterns. For instance, Huang et al. [29] proposed the use of a probabilistic model, which is the Latent Dirichlet Allocation, that aims at identifying patterns based on treatment activities and corresponding time stamps in order to reveal the temporal structure of the discovered patterns. Michalowski et al. [33] used a Bayesian Belief Network to model the radical prostatectomy clinical pathway, focusing on patients length of stay, categorizing it as ”met” or ”delayed” given the patients’ activity and outcome. Lin et al. [32] resort to Hidden Markov Models to uncover clinical pathways.

A HMM is a stochastic probabilistic model to mold time-series data, allowing for easy incorporation of new instances to update the model. These models have proven to accurately representing clinical pathways in the past. By identifying events of interest within healthcare data, for instance, medication swaps, treatment steps, diagnostic steps, medical consult activity, it is possible to build Markov Chains that help uncover the most probable pathways within a certain population. In these models, a certain state or event generates an output state, or in other words, transitions to a next state according to transition distributions. This represents a simple was of uncovering most prevalent patterns when considering a certain set of patients with certain conditions, when trying to evaluate a certain activity within the medical care process.

2.6 Clustering

Clustering can be defined as the recognition of natural groups within multidimensional data based on a certain similarity measure, carrying an important role in pattern recognition and machine learning techniques [34]. Resorting to cluster analysis, it is possible to group together certain objects that admit similar characteristics, forming homogeneous partitions, while aiming at maximizing heterogeneity amongst distinct groups. Hence, the goal is to maximize intra-cluster similarity and minimize it inter-class wise, allowing to understand which entities are closer to each other, as well as the most distinguishable ones, which is useful when aiming at identifying sub groups within large data sets [35].

In the context of healthcare data, clustering has demonstrated to be a powerful tool for determining certain patterns in medical datasets [36]. Given the amount of data generated in the healthcare systems, it is advantageous to find similar structures of data in order to detect certain patterns and similarities amongst patients. For instance, it is interesting to divide patients that suffer from a same disease into subgroups, so to identify different characteristics of the disease and its symptoms and evolution, allowing health professionals to have a more personalized treatment depending on how a patient is classified.

2.6.1 Clustering techniques

Since the clustering process is based on grouping elements that are similar to each other, in order to be able to distinguish several groups of elements from one another, it is necessary to base the method in a similarity measure. The most common way of evaluating how similar a pair of elements is, is resorting to distance metrics for instance, the Euclidean distance. For each pair of elements in a data set, it is necessary to use a similarity measure, from which the clustering technique itself is based, grouping elements close to each other, distancing them from the ones that are more distinct.

There are several clustering techniques that one can use, depending on the goals and data available for the process. Two main types of clustering stand out in the literature, namely hierarchical and partitional clustering [34] [36]. Hierarchical clustering groups data with a sequence of nested partitions [36], creating a tree-like structure, known as a dendogram, admitting two distinct strategies to do so, splitting the data hierarchically, resorting to a ”bottom up” approach, which is called agglomerative clustering or to a ”top down” approach, known as divisive clustering. The first one starts off by considering each observation as an individual cluster, merging the two most similar clusters successfully until all objects form one big cluster. Divisive clustering, in turn, begins by considering all elements in one cluster and splitting them until each object cluster is formed by only one object. On the other hand, partitional clustering divides objects directly into a certain amount of sub-groups without resorting to a hierarchical architecture, aiming at minimizing a certain criteria, for instance, a square error function [36]. A popular example of this type of clustering approach is the k-means algorithm, which has as main goal minimizing the intra-cluster distance. It starts with k centroids representing k distinct clusters, assigning each observation to the closest one, recalculating the centroids as the method evolves, until an optimal convergence is reached.

2.6.2 Clustering Validation

One of the main challenges associated to the clustering process is related to the choice of the most appropriate number of partitions that best fit the data set in hands. A clustering algorithm groups together elements that are close to each other based on a certain measure, while maximizing the inter-cluster distance, however, the problem is in choosing how many distinct groups to choose so that these distances are optimum. Cluster validation is the process of choosing the optimum number of clusters that best fit the data.

There are several clustering indices that can be used to infer on the partitions stability. The literature points out three types of measures that can be used to evaluate cluster stability, which are internal, external and relative measures. The first ones use features that are intrinsic to the data, external ones are based on reference, meaning that the clustering results are compared with labels that already exist in order to infer on the quality of the clusters, and finally, relative measures are used to compare different groups of clusters, generally obtained through the same algorithm, but using distinct features or parameters. Generally, internal measures are the most used since clustering is a technique mostly used in unsupervised learning, when there are no labels available. These measures are used to reckon upon a clusters’ compactness and separation of data, meaning that one cluster should be as compact as possible, being formed by elements close to each other, while the separation indicates that data from different clusters should be as diverged as possible.

The silhouette score (SS) is an internal measure widely used for this purpose, since it takes into consideration both compactness and separation. It analyses both the distance of each element to the elements belonging to the same cluster and to the elements that form the closest neighbour cluster. It is a versatile measure, since it can be used to evaluate a single element, a cluster itself or all clusters obtained from a certain clustering algorithm. This measure for one single data point *i* is inferred as follows:

$$\mathrm{SS}(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.1}$$

where *a(i)* is the distance of *i* to the elements of its own cluster, given by the average of individual distances and *b(i)* represents the distance to the closest cluster, given once again by the average distance of *i* to all data points in that neighbour. From these silhouette scores, which range between -1 and 1, it is possible to evaluate how well a certain data point is inserted in its cluster, compared to how far it is from others. When the inter-cluster distance (*a(i)*) is much smaller than the intra-cluster distance (*b(i)*), an SS close to 1 will be obtained, indicating that the data point in question is well clustered. On the other hand, if *a(i)* is much higher than *b(i)*, an SS approximating -1 is obtained, meaning that *i* is closer to elements from a different cluster in opposition to its own cluster neighbours, which indicates that this specific data point is badly classified. The SS for a whole cluster is given by the average silhouette values across all data belonging to that subset.

Another internal measure is the Dunn index, which also considers inter- and intra- cluster distances and is calculated by dividing the minimum inter-cluster distance by the maximum intra-cluster distance. This index is very similar to the silhouette one, since it also identifies clusters that are compact, being formed by members that are close to each other and widely separated from neighbour clusters. Other internal measures include the Calinski-Harabasz, Davies-Bouldin and Ray-Turi, which all have the same goal of an internal measure.

External indices include, for instance, the Rand Index, which provides a measure of similarity assessment between two partitions, comparing two clustering processes, which is useful to decide on the best option for the data set in hands.

2.6.3 State of the Art methods for temporal sequence alignment and patient clustering

In order to study and better understand multimorbidity patients and their disease patterns, it is crucial to consider time information, apart from the associations between the diseases itself. This study has its focus on two main components, the one of pairwise sequence alignment, focusing on clinical data's temporal information, as well as the process of clustering patients itself.

Giannoula et al. [37] published a population based study that aims at identifying temporal patterns in patient disease trajectories. To that end, firstly, pairwise disease associations as well as their temporal directionality are assessed, moving on to a Dynamic Time Warping (DTW) based clustering algorithm. This DTW technique is applied on the common disease trajectories in order to group them according to shared temporal patterns. In other words, the authors resource to a novel unsupervised clustering algorithm which is based on DTW, that aims at grouping the trajectories with similar diagnoses patterns into clusters. This technique works by determining an optimal path between two sequences considering certain rules and restrictions, that minimize the total distance between them. Posteriorly, the sequences can be warped (aligned) non-linearly along the time dimension, according to the optimal warping path. These distances are then used for the clustering algorithm: every new disease trajectory in analysis is compared, through mean distances, to all disease trajectories belonging to each already existing clusters. Hence, the mean distance (or global cost) between the new trajectory and the members of existing clusters are assessed and the minimum value of that cost is identified, assigning the new trajectory to the respective cluster.

The Needleman-Wunsch (NW) algorithm was one of the first methods used to align and compare protein and nucleotide sequences, resorting to insertion of gaps rather than warping as the DTW algorithm [38]. Given two temporal sequences of medical events, this algorithm calculates an accumulated score matrix considering a standardized gap penalty for when it is necessary to insert a gap and a defined scoring system that denotes similarity between the two sequences. Furthermore, the NW algorithm is able to identify optimal alignment paths by tracing back from the accumulated score matrix, maximizing these scores along the path.

The Smith-Waterman algorithm is another example for the sequence alignment purpose, which is very similar to the NW one, however, it is a local sequence alignment algorithm, while the NW one is global [38]. The main difference from NW is the non-existence of negative scores, since these are set to 0. The optimum local alignment path is done in the same way, by starting at the element with the highest accumulated score. From there, the local alignment path with the highest similarity is identified by tracing back in the matrix choosing to go through the path that leads to better accumulated scores, until a zero is met.

AliClu algorithm

With the sole purpose of developing a method that resorted to longitudinal data available from EMRs in order to stratify patients, considering as well temporal information, an algorithm named AliClu was developed [39]. AliClu is a Python implemented algorithm, from 2018, used to perform clustering of temporal data, through combination of the Temporal Needleman-Wunsch (TNW) algorithm and hierarchical clustering. When implemented, this algorithm was tested both in synthetic data sets, as well as with data from the Reuma.pt databse, particularly to study therapy switches regarding biological drugs administered to rheumatologic patients, considering as well the time elapsed between switches.

The AliClu algorithm itself was implemented together with a pre processing script, which receives the events of interest in a panel data format, delivering a set of temporal sequences associated to the patients involved. These temporal sequences, referred to as prefix-encoded (PE) sequences, represent a progression of events over time and are represented as follows:

$$0.A, t1.B, t2.C, ...,$$

where letters represent the events under study and numbers between consecutive events represent the time elapsed between them. These temporal sequences are then given as input to the algorithm, together with the patients ID's, which outputs the more appropriate set of clusters, considering the choice of parameters.

AliClu has incorporated a method of learning temporal patterns from patient data resorting to an algorithm that aligns pairs of temporal sequences and uses obtained alignment scores in order to proceed to the process of hierarchical clustering. The algorithm used to complete this alignment is the Temporal Needleman-Wunsch. Extended from the original NW algorithm, the TNW, besides considering the matching between sequence events, also takes into account time variables between consecutive events, during the alignment process [39]. When studying data describing certain states, along with their corresponding duration, this method is very useful, allowing to encounter similarities between patients based on their medical histories [40].

That being said, after obtaining the PE sequence for each patient, it is possible to align all patient pairs using the TNW algorithm. An optimal alignment is influenced by the right choice of a user defined set of parameters, namely an appropriate scoring schema, gap penalty and temporal penalty. The scoring scheme used in this work, just as Rama et al. [40] [39] is a simple one, which provides a score of 1 to exact-matching events and -1.1 to mismatches. The gap penalty is assigned when the events mismatch, so that events that do not match are assigned a score of -1.1 and a gap penalty of a certain value, normally, between -1 and 1. Designating a score of -1.1 to non-matching events reflects the preference of inserting a gap in each sequence, instead of aligning events that do not match. In addition to these parameters, also necessary for the NW algorithm, the TNW resorts to a temporal function, in order to consider the time variable between events. This function simply uses the percentage discrepancy between event transitions in comparison to obtain a time penalty and requires a defined maximum value that can be assigned to temporal differences [40].

Considering then all these parameters, a pairwise alignment is carried out between all pairs of sequences, obtaining a final score for each of the alignments. These scores are summarized into an N

x N similarity matrix, N representing the number of patients in the dataset. Each entrance of the matrix represents the alignment score of the PE sequences between those two patients. However, for the hierarchical clustering, it is only necessary to compute the upper triangular part of the matrix, since a similarity matrix of this kind is diagonal and symmetric, with the highest values in the diagonal, as they correspond to the similarity scores between a sequence and itself.

The AliClu algorithm, in combination with the sequence alignment algorithm, resorts to hierarchical agglomerative clustering to obtain the most appropriate patient stratification. To this end, it is necessary to have a distance matrix containing the distances between pairs of sequences, since the process of hierarchical clustering involves merging the clusters that are closer to each other (shorter distance), which are the ones with highest similarity score. Hence, prior to the hierarchical clustering itself, it is necessary to convert the similarity matrix obtained from the TNW algorithm to a distance matrix. This is simply done by inverting these scores, followed by a resampling, so that distance scores are all higher or equal to zero [40]. In the case of this distance matrix, it is diagonal and symmetric, just as the similarity matrix, however, the values in the diagonal are zero, representing the distance between a sequence and itself.

Furthermore, this method of agglomerative clustering demands for a distance metric, known as linkage function, that evaluates which clusters are closer or further away from each other, in order to know which clusters to successively merge. There are several linkage functions available to define how the dissimilarity between objects is measured, namely single link, complete link, average link, centroid link as well as Ward's method, being the latter the chosen method to work with in this analysis. Ward's method measures the distance between two clusters by assessing how much the sum squares will increase when merged, taking into consideration distances between all cluster objects and their centroid, choosing the successive clustering steps so that the increase of the error sum of squares is minimized [41]. The results of hierarchical clustering are normally depicted in a dendogram, where it is possible to visualize the successive merges, as well as the distances involved.

In addition to the hierarchical clustering process itself, AliClu was implemented with an automatic bootstrap approach for validation, due to the fact that these clustering methods do not explicitly set the optimum number of clusters [39]. To sum up, the bootstraping approach consists of, posterior to obtaining the distance matrix for the complete data set, dividing the data M times, where M is the user-defined number of bootstrap samples, using three quarters of the patients to create a new distance matrix and performing agglomerative clustering on that bootstrap sample. This process is repeated M times. Every time the agglomerative clustering is performed, five clustering indices, which are briefly described below, are calculated between the partition resulting from the clustering with all objects and the partition of the clustering with three quarters of randomly sampled elements.

By varying the parameters necessary to run AliClu, different clusters are obtained, which need to undergo a qualitative and quantitative evaluation. AliClu already has implemented 5 clustering indices, specifically Rand Index, Adjusted Rand Index, Fowlkes and Mallows, Jaccard Index and Wallace's coefficients, that help assess how good the obtained clusters turned out to be. All these indices represent measures of similarity between two distinct clusterings, calculated in different ways. The goal is to obtain these values to compare the clusters obtained considering all objects and the one considering three quarters of randomly sampled objects, in order to perform bootstrapping validation. For each number of clusters analysed, the maximum values of the indices within the indices results obtained with the bootstrap samples are set as the final values for these clustering indices, in order to allow for a posterior analysis of what the optimum number of clusters is for the data set in hands. Resulting from this bootstrap analysis, a table is presented with the values of the referred indices for each *k* considered, which is delivered by the algorithm as a pdf file. The final decision about the best number of clusters to choose lays on choosing the number of clusters, *k*, that yield the maximum index values, however, not all indices will have their maximum in the same *k*. Hence, the final decision lays on a compromise decision between the values of these indices and the standard deviation obtained for each *k*, also presented in the final pdf.

Considering all this information described above, it is possible to run AliClu in two distinct manners: automatic and semi-automatic. With the automatic version, it is the algorithm itself who makes a decision about the optimum number of clusters, based on the considered indexes calculated. On the other hand, the semi-automatic version allows the user to analyse the clustering indexes and dendograms obtained for all sets of parameters analysed and making a final decision on the most appropriate number of clusters.

Finally, it is also important to infer on the clusters stability. AliClu returns a pdf file with this analysis, which is done after the number of clusters have been chosen. To measure the stability of each of the final clusters obtained, three measures of correspondence are used, which are Jaccard, Rate of Recovery and Dice coefficient measures. The same bootstrap method is applied here in order to compare two clusters, one of them resulting from the clustering of a randomly drawn bootstrap sample from the original data set. The mean, median and standard deviation of these indices are calculated and the goal is to look for higher (closer to 1) average values and lower standard deviations (closer to 0), since it is difficult to define a threshold to consider a cluster as being stable.

3

Methodology

Contents

The main focus of this work was to identify patterns within a subset of multimorbidity patients who suffer from Dementia, which had activity in Hospital da Luz Lisboa (HLL) from January 2007 to August 2021. We completed an initial phenotype screening and characterisation of the study cohort, followed by a clinical pathway analysis resorting to two distinct methods: Hidden Markov Models and AliClu clustering algorithm. A more detailed description of the methods used is presented in the next sections.

3.1 Available data and initial phenotyping

The available data, necessary to develop this work, is handed in a panel data format. Panel data, or longitudinal data, document the same sample throughout different points in time, hence, each sample is distributed across several lines in the input data. One of this work's main goal is to study patient activity within the hospital, thus, the time points considered represent consults of a certain medical specialty attended. That being said, each patient is distributed across several lines in the available data, where each line represents a consult attended, along with the date of occurrence. It is important to point out that each patient is identified by an ID, which is simply a cardinal number assigned randomly to the data set in question, allowing for total data protection, being that the real identifications are confidential.

The concern of this work is in patients with MM, more specifically, in patients attending General and Family Medicine (GFM) consults, which is a medical specialty with some heterogeneity, hence the importance of grouping the patients, taking into account their activity throughout different specialties.

Furthermore, in order to specialize this analysis, it was considered advantageous to focus the study on patients suffering from a certain common disease, so to understand how disperse are the pathways related to that specific illness. A phenotypic heterogeneity regarding patients with Dementia has been previously flagged, mainly due to the fact that this disease can result from a number of distinct factors and other conditions, with diverse etiology and pathophysiology. Within patients with Dementia, various symptoms, disease trajectories, as well as individual onset ages can be identified. Given all these challenges and the importance of better understanding these patients, we selected them as our study cohort. Patients suffering from Dementia were then isolated and handed in the mentioned format for the process.

Within the dataset used, there were twenty five different medical specialties considered, listed in Table 3.1, with the respective event label assigned. An initial assessment of the prevalence of the different types of consult in the whole data set, as well as separately in the female and male data sets was carried out.

Moreover, in order to further understand the data set in hands and its heterogeneity, a preliminary evaluation of the population as a whole, as well as by gender was done. Hence, a phenotype screening was carried out regarding Dementia patients, prior to moving on to more specific analysis. To this end,

Table 3.1: Medical specialties considered for patient activity analysis, along with the corresponding event label assigned by the algorithm.

Medical Specialty	Label assigned
Nutrition and Dietetics	A
Hematology	B
Ophthalmology	C
Cardiology	D
Nephrology	E
Orthopedics	F
Anesthesiology	G
Urology	H
Dermatology	I
Obstetrics and Gynaecology (Ob-Gyn)	J
General and Family Medicine (GFM)	K
Immunoallergology	L
Otorhinolaryngology (ORL)	M
Internal Medicine	N
Endocrionology	O
Pneumology	P
Gastroenterology	Q
Physical Medicine and Rehabilitation	R
Oncology	S
Dental	T
Psychiatry	U
Rheumatology	V
Neurology	W
Neurosurgery	X
Surgery	Y

data on patients age, chronic diseases and consults attended was gathered in order to observe the distributions corresponding to these characteristics. Furthermore, so to understand the prevalence of the chronic diseases suffered by these patients, as well as their co-occurrence in the data set, graphs of co-occurrence were obtained, for the population as a whole, as well as separately for female and male patients.

Following steps are mainly centered on hospital activity, in order to identify recurring patterns of medical consults within Dementia patients.

3.2 Using Markov Chains to identify activity patterns in Dementia patients

With the goal of obtaining an overall view and understanding of the activity of patients with Dementia regarding medical consults, through estimation of Markov chains, it was possible to determine the most prevalent transitions between consults. This was accomplished by formulating a transition matrix (TM), which shows the transition probabilities between two states (*i.e.*, consults). Given a square matrix of all possible medical specialities, we can calculate the conditional probabilities of moving to a second speciality consult (*j*), given the previous one (*i*). This is achieved by dividing the number of times that each transition occurs by the prevalence of the origin medical speciality consult, filling the TM as follows:

$$TM_{ij} = P(i \rightarrow j) = \frac{\#(i \rightarrow j)}{\#i} \tag{3.1}$$

Two different approaches were embraced to estimate these conditional probabilities, one of them considering consecutive appointments between the same medical speciality, and a second one treating these as one, in order to better identify consultation patterns without considering follow-up consults in the same speciality.

3.3 Using the AliClu algorithm to stratify Dementia patients

In order to stratify Dementia patients considering their activity patterns within the hospital, the AliClu algorithm was used. However, slight adaptations considering the data in hands, which will be explained further ahead, were necessary, both on the algorithm itself, as well as in the pre processing step.

In summary, first of all, the available data from HLL concerning consult activity of patients with Dementia was converted from panel data format into the appropriate sequences to be used as input for AliClu. Then, an optimization process was implemented in order to reach the best stratification possible and lastly, the final clusters were obtained.

3.3.1 Data Pre-Processing

As previously mentioned, the AliClu algorithm is available alongside a script that takes as input the data in panel format, outputing the required PE sequences which will serve as input to the clustering algorithm itself. This script, first of all, assigns a label to each event, and depending on the time point format of the raw data, two different pre processing steps are available. In this case, the time point of the events were represented as dates. Resorting to the patient ID, his sequence of consults attended and their occurrence date, this script delivers the temporal sequence for each patient in the data set. Adjustments to this pre-processing script were necessary in order to adapt it to the cohort under study, namely, an appropriate dictionary had to be defined considering the amount of consults being handled. It is important to point out that when an event is repeated consecutively, the algorithm counts as only one event, merging the elapsed times into one.

Furthermore, it was necessary to implement an additional pre-processing script in order to filter some patients who did not meet the appropriate conditions to be served as input to the AliClu algorithm.

Specifically, in view of the fact that the focus is on patients with MM, patients who only have one medical consult present in their temporal sequence, removing these patients is an intuitive step. Moreover, in order to avoid *outliers* as much as possible, an additional filtration is done, removing all patients who attend a number of different consults superior to a certain threshold. This threshold is given by the 95 percentile regarding the number of consults attended throughout a patient's pathway.

Figure 3.1 represents the overview of the necessary steps to achieve the desired PE sequences.

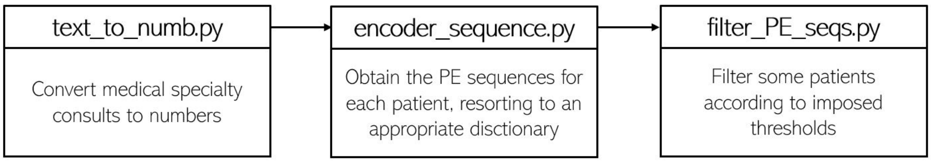

text_to_numb.py
Convert medical specialty consults to numbers
encoder_sequence.py
Obtain the PE sequences for each patient, resorting to an appropriate disctionary
filter_PE_seqs.py
Filter some patients according to imposed thresholds

Figure 3.1: Overview of steps necessary for the data pre-processing.

Subsequently to this step, the patients who meet the appropriate criteria to undergo the clustering process, are characterized only by their ID and respective temporal sequence, which provide the patient's clinical history regarding consult activity.

Summarizing, an example of the input data for the AliClu algorithm is represented in Table 3.2. Letters represent the consults attended and the time stamp between each event pair represents the time elapsed between them, in days.

Table 3.2: Example of the temporal sequences corresponding to two different patients, representing the AliClu input data format.

id_patient	aux_encode
1	0.A,t1.B,t2.C
2	0.D,t3.E,t4.B,t5.F

3.3.2 Clustering Indices

Besides the clustering indices considered in the AliClu algorithm, mentioned in Section 2.6.3.A, in order to get a more visual analysis of the clusters content, another index was added to the process. For a set of clusters obtained, the silhouette score (SS) is computed, both for each individual element, as well as the average SS across all samples and per cluster. A silhouette score measures how close an element is to it's own cluster, compared to how close it is to others [42]. These scores are then shown in a plot where it is possible to analyse in a more intuitive manner which clusters seem to contain more similar objects and which ones do not, for example by checking if the average cluster silhouette scores is¡ below or over the overall SS.

3.3.3 Parameter Optimization

AliClu was designed to return one set of clusters for the best combination of gap penalty, g, and number of clusters, k, being the temporal penalty, Tp, established in the beginning of the process and hence, not iterated during the development. AliClu was designed to run in an automatic and semi-automatic manner, both returning merely one set of clusters, depending on the considered optimum parameters:

- **Automatic mode:** the algorithm itself decides which is the optimum number of clusters and gap penalty which lead to the best clusters. The values between which the choice is made are previously user defined and the decision is based on the results of the five clustering indices mentioned in the previous section.
- **Semi-automatic mode:** involves an interruption of the process, so that, through analysis of the pdf containing the values of the cluster indices for each gap penalty and number of clusters, the user decides which gap and number of clusters gives the best results.

The AliClu algorithm is a very sensitive one when it comes to the choice of parameters, namely gap penalty and number of clusters, but also temporal penalty, meaning that a change in these parameters will deliver completely different clusters, which underlines the importance of tuning these parameters in the most efficient way possible.

That being said, in order to optimize these three key parameters in the most efficient way possible, a more intensive search, based on the average SS, for the best set of (g, Tp, k) was implemented. So, AliClu was adapted to return a set of clusters for each combination of the three parameters. For each set of clusters, the average SS was computed, since this metric is a good indicator of the cluster's content. From the collection of average silhouette scores obtained for each set of clusters, the optimum parameters were chosen by searching for the highest score obtained.

For the purpose of searching through the broader range of parameters possible, considering the data set in hands, an initial implementation was carried out, serving the sole purpose of deciding on the best set of parameters. Hence, the whole dataset was split into six partitions of equal size and an adaptation of the original AliClu algorithm was applied to those partitions. This implementation returns, for each partition, the average SS obtained considering each set of parameters (g, Tp, k) analysed. Algorithm 1 details the steps of this modified AliClu, silhouette score based, parameter optimization process. It is important to point out the fact that the number of partitions was chosen based on the size of the data set in question, so that the resulting partitions size would be large enough to deliver appropriate clusters and small enough to make the running process the least heavy and complex as possible.

The choice of parameter values to be analysed is just as crucial as making a final decision about these. Due to the fact that, generally, for gap values higher than 0.5, considering the scoring system in

Algorithm 1 Parameter optimization process

```
1. Define range of $g$ and $Tp$ to analyse.
2. Partition the data into $n$ random equal-sized subsets.
3. For each partition:
   • For each $g$:
     – Initialize .csv file to keep track of average silhouette scores for clusters resulting from
       gap penalty = $g$.
     ∗ For each $Tp$:
       – Perform pairwise sequence alignment;
       – Convert similarity matrix to distance matrix;
       – Perform hierarchical clustering;
       – Initialize silhouette average list for corresponding $Tp$.
         □ For each $k$:
         – Cut dendogram according to $k$;
         – Print clusters found into .csv files;
         – Calculate silhouette scores for the $k$ clusters found;
4. Visualize evolution of average silhouette scores considering the parameters analyse.
5. Decide on optimum parameters ($g$, $Tp$, $k$) considering the maximum average silhouette score
   reached and analysis of .csv files containing the clusters obtained.
```

place (1 for matches, -1.1 for mismatches), the algorithm tends to align mismatch events, instead of inserting gaps where the sequences don't match, while for gap values lower than -0.5, total misalignments may occur (e.g., the algorithm may not even align the events that actually match) [40], the values of gap penalty chosen to analyse in this work were only comprised between -0.5 and 0.5.

Regarding the temporal penalty, when using very low values (Tp $<$1), the obtained alignments are usually the ones that would be obtained if ignoring temporal variables. On the other hand, for temporal penalties close to 10, once again, total misalignments may occur [40]. Hereupon, values of 1, 2, 5, 7 and 10 were tested for temporal penalty. With respect to the minimum and maximum number of clusters analysed, these were set to 2 and 20, respectively.

Keeping these restrictions in mind, for a first analysis, it was hypothesized that the mentioned gap and temporal penalties for the TNW algorithm, as well as the Ward linkage function for the hierarchical clustering and the number of bootstrap samples set to 250 for the validation process, were indeed the better choices when compared to other values. Nonetheless, so as to not take these assumptions for granted, we did a second analysis to test other variables and guarantee that better results weren't indeed reached.

Firstly, using the optimum values of g and Tp reached, the linkage function used in the hierarchical clustering process was modified in order to test the single, complete and average linkage methods and the results of the average SS obtained for each partition using each one of the four methods (including ward's) were compared. The next step was in order to guarantee that the number of bootstrap samples

(*M*) was appropriately set. To this end, *M* was increased so to analyse the evolution of the average silhouette scores with a growing number of bootstrap samples. Subsequently to assuring that Ward's linkage method was indeed the best choice and a number of bootstrap samples set to 250 was enough for the data set under study, the next step focused on testing gap penalties outside the interval which was considered better. Hence, values ranging from -1 to -0.6 and 0.6 to 1 were tested for this parameter, analysing once again the average SS across all partitions. It is important to point out that, to support this SS analysis, the obtained clusters in each optimization step were visually analysed so to guarantee that the changes in the SS results went accordingly to what was observed.

3.3.4 Clustering of female and male data sets

With the intention of grouping patients and identifying patterns considering a more targeted approach, the process described above was repeated for the female and male data sets alone. Firstly, the pre processing filtering step was set to remove those patients who did not fit the most appropriate criteria to be included in this stratification process.

Since the parameter choice is crucial and dependent on the data set, once again, the optimization process was repeated, in order to search for the best temporal and gap penalty values for each one of the patient groups. Regarding the optimum number of clusters, once again, this parameter was not chosen based on the parameter validation process, since the optimum number of clusters can vary when dealing with different sized data sets and the goal is optimizing for the whole male and female subsets. The criteria for choosing *k* which would lead to the optimum number of clusters was once again the average SS across all samples.

3.4 Cluster analysis and phenotyping

Subsequently to finding the optimum set of parameters which lead to the best patient clusters, it is imperative to assess their quantitative and qualitative reliability. To this end, through every phase of the parameter optimization process, a manual observation of the so considered optimum groups obtained was done, in order to confirm that the algorithm was indeed grouping patients which presented a similar pathway regarding consult attendance.

Moreover, besides the average silhouette scores calculated across all samples, the mean silhouette was also calculated within each cluster, so to assess the similarity of the elements within a certain cluster, when compared to elements allocated to different groups.

AliClu has a cluster stability analysis process integrated where, for each of the final clusters, three metrics of three different indices are calculated, which are the median, average and standard deviation of the Jaccard Index, Dice Coefficient and Recovery Rate. By analysing these metrics, it is possible to conclude about the quality and stability of each developed cluster, considering that the average values should be as close to 1 as possible, while the standard deviations should be closer to 0.

Subsequently to the quantitative analysis of the obtained clusters, we repeated the phenotype screening and characterisation process for each of the subsets obtained, in order to detect possible patterns that may relate the prevalent medical consult of each one of them with other phenotypes or characteristics. This included exploring age, gender, number and type of chronic diseases, medications and hospital admissions of patients within each cluster.

Regarding the gender distribution per cluster, Fisher's exact test was used to assess the relevance of the gender proportions found relative to the whole data set proportion. This test is used to determine if a significant association between two categorical variables in a contingency table exists [43]. To this end, a contingency table is generated for each cluster, where the categorical variables are gender and whether or not a patient belongs to it, as exemplified in Table 3.3. *a* and *b* represent the number of female and male patients, respectively, which belong to the subgroup in question. *f* and *m* represent the total number of female and male patients that form the cohort. Hence, *f* - *a* and *m* - *b* represent the amount of female and male patients that do not belong to the cluster in question.

Table 3.3: Example of a contingency table used for Fisher's exact test.

	Females	**Males**
In cluster	a	b
Not in cluster	f - a	m - b

Fisher's exact test allows to determine the probability that a gender proportion found in a certain cluster, as or more extreme than the whole data set proportion, is caused by random chance. With the information given by the contingency tables, it is possible to assess the p-value. A p-value is only significant when it is lower than a threshold of 0.05. Having a p-value lower than 0.05 associated to a certain cluster, we can conclude that the gender proportion of that subgroup is not due to the original proportion.

3.5 Medication analysis

Aiming at finding associations between the medication intake of the considered group of patients and the comorbidities and prevailing medical speciality consults within each cluster, we pursued a medication analysis per cluster obtained. First of all, we identified patients who did not have any prescriptions associated to their process and calculated the average number of prescriptions per patient, as well as the number of different medicines prescribed per patient, both for the whole data set and per cluster obtained.

The next step was identifying the most prevailing medication within the cohort and each cluster.

However, due to the vast amount of medicines identified in the dataset and knowing that several ones have the same purpose, we grouped the medication present in the cohort by class. The most relevant medication classes regarding the patients under study were identified and the different medicines were assigned to them, having performed an analysis not of a medicine itself, but of each of the prevailing classes of medication, both for the whole dataset and for each individual cluster. Table 3.4 presents the chosen classes, as well as the medicines that are part of each one.

3.6 Hospital admission and emergency analysis

In order to understand the tendency of these patients having emergency episodes and the need to be admitted to the hospital, we did an analysis of these occurrences within the considered time window. First of all, we did a survey of the fraction of patients with at least one hospital admission or emergency episode since January 2007 until August 2021. Then, the average number of each occurrence per patient was also assessed. Furthermore, in order to understand the amount of hospitalizations and emergencies and how they evolved throughout the years, the percentage of admitted patients was determined for each event. Finally, an analysis was carried out per cluster, aiming at comparing the stronger or weaker tendency of each subgroup being admitted or having an emergency. To this end, an Odds Ratio (OR) analysis was carried out.

An OR represents a measure of association between an exposure and an outcome, meaning, it gives the odds of an outcome occurring given a certain exposure, when compared to the odds of occurring when that exposure does not exist [44]. The value of the OR is defined as follows:

- OR = 1: means that the exposure does not influence the outcome;
- OR $<$1: the exposure causes lower odds of the outcome occurring;
- OR $>$1: the exposure leads to higher outcome odds.

In our specific case, the goal is to understand if a patient belonging to a specific cluster (exposure) is more or less probable of being admitted or having an emergency episode (outcome).

The formula for calculating an OR is the following:

$$\text{OR} = \frac{a/c}{b/d} \tag{3.2}$$

where *a* represents the number of exposed cases, *b* the number of exposed non-cases, *c* the number of unexposed cases and *d* the number of unexposed non-cases. In this specific case, aiming at calculating the odds ratio for a patient in a certain cluster being admitted to the hospital or having an emergency episode, *a* corresponds to the number of patients in cluster *C* that had an episode, *b* represents the

Table 3.4: Considered medication classes and respective included medicines.

ACE / ARBs	Benazepril, Fosinopril, Lisinopril, Captopril, Enalapril, Ramipril, Moexipril, Quinapril, Trandolapril, Candesartan, Eprosartan, Irbesartan, Losartan, Olmesartan, Telmisartan, Perindopril, Valsartan, Cilazapril, Sacubitril + Valsartan
Analgesics	Paracetamol, Magnesium metamizole, Paracetamol + Thiocolchicoside, Buprenorphine, Fentanil, Morphine, Oxycodone + Naloxone, Tapentadol, Tramadol
Ansiolitics	Bromazepam, Estazolam, Pregabaline
Anticoagulants	Apixaban, Rivaroxaban, Dabigatran, Edoxaban, Dabigatran
Antidepressives	Citalopram, Escitalopram, Fluoxetine, Fluvoxamine, Paroxetine, Sertraline, Vortioxetine, Vilazodone, Trazodone, Mirtazapine, Duloxetine
Antidiabetics	Empagliflozine, Canagliflozine, Dapagliflozine, Metformine, Insulin glargine, Insulin lispro, Insulin detemir, Insulina asparte, Human Insulin, Insulin degludec
Antiplatelets	Acetylsalicylic acid, Clopidogrel, Ticagrelor
Antipsychotic	Olanzapine, Quetiapine, Risperidone
Benzodiazepines	Alprazolam, Chlordiazepoxide, Clonazepam, Clorazepate, Diazepam, Estazolam, Flurazepam, Lorazepam, Midazolam, Oxazepam, Temazepam, Triazolam, Quazepam
Beta-blockers	Atenolol, Atenolol, Bisoprolol, Betaxolol, Bisoprolol, Betoprolol, Carvedilol, Celiprolol, Labetalol, Penbutolol, Timolol, Carteolol, Metoprolol, Nadolol, Nebivolol, Pindolol, Propanolol, Sotalol, Acebutolol, Inderal
Bronchodilators	Salbutamol, Salmeterol, Formoterol, Vilanterol, Ipratropium bromide, Tiotropium bromide, Aclidinium bromide, Glycopyrronium bromide, Theophylline
Calcium channel blockers	Amlodipine, Nifedipine, Diltiazem, Felodipine, Verapamil, Lercanidipine
Dementia (Cholinesterase inhibitors + NMDA receptor antagonist)	Donepezil, Galantamine, Rivastigmine, Memantine
Diuretics	Furosemide, Torasemide, Hydrochlorothiazide, Indapamide, Chlorothiazide, Metolazone, Amiloride, Spironolactone, Triamterene
Non-steroidal anti-inflammatory drugs (NSAIDs)	Diclofenac, Clonixin, Celecoxib, Etodolac, Ibuprofen, Indometacin, Ketoprofen, Naproxen, Piroxicam
Statins	Atorvastatin, Rosuvastatin, Sinvastatin, Pravastatin, Lovastatin, Fluvastatin, Pitavastatin, Sinvastatin, Pravastatin, Lovastatin, Fluvastatin, Pitavastatin, Atorvastatin, Rosuvastatin
Steroids	Prednisolone, Betametasone, Dexametasone, Hidrocortisone, Metilprednisolone, Deflazacort
Vasodilators	Nifedipine, Magnesium Sulfate, Cinnarizine, Seloken, Monocordil, Vasativ, Isordil, Unoprost, Cialis, Atenol, Vertizine D, Vicog, Levitra, Ginkgo, Cebralat, Sustrate, Caltren, Claudic, Nitroglicerine
Vitamins	Bioflavonides, Folic acid
Thyroid medication	Levothyroxine, Propylthiouracil

number of patients in the same *C* that did not have an episode, *c* is the number of patients not in *C* that had an episode and *d* the number of patients not in *C* that did not have an episode.

4

Results and Discussion

Contents

In this section, all the results concerning the methods used throughout this work are presented. The first part includes the product of the initial phenotype screening concerning patients with Dementia. Subsequently, moving on to the clinical pathway analysis, the outcomes regarding the Markov Chain approach, as well as the ones concerning the application of the AliClu algorithm are presented.

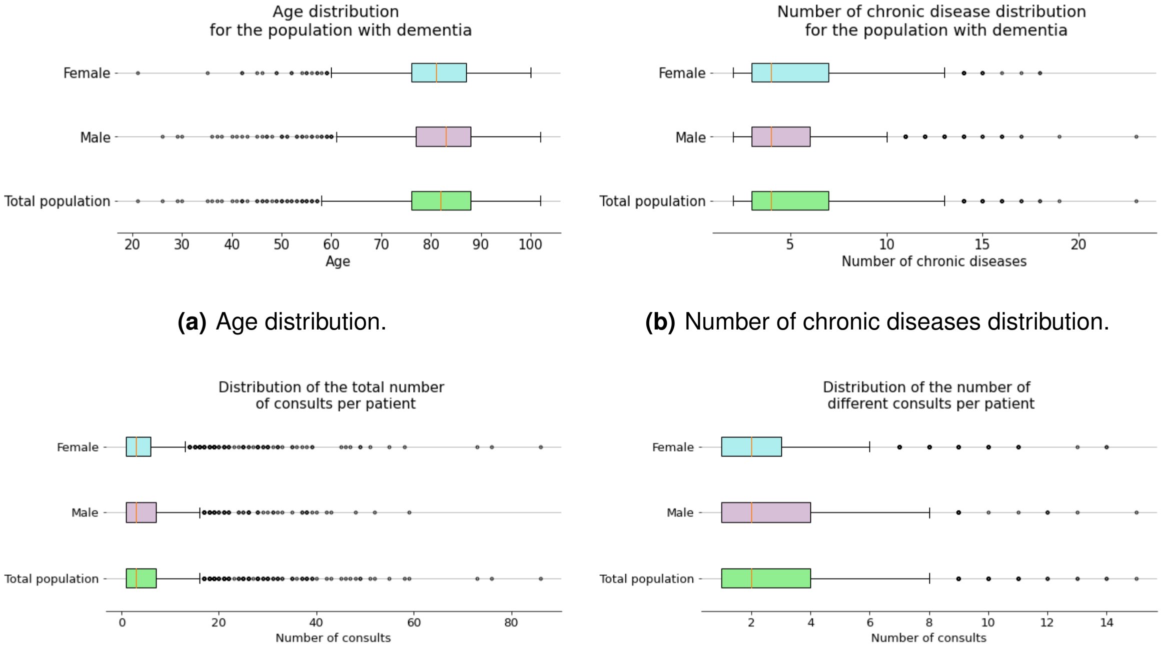
4.1 Available data and initial phenotype screening of Dementia patients

As formerly mentioned, this work focused on patients suffering from Dementia. From the 54 827 multimorbidity patients identified in Hospital da Luz Lisboa between January 2007 and August 2021, through ICD-9 codes and specific keywords related to Dementia, it was possible to identify those MM patients who suffered from this disease. A total of 1 924 patients with Dementia, where 1 147 are female and 777 are male, were identified, together with the information on 20 033 consults attended relative to these patients during the time frame considered. In order to better understand the cohort under study, we did an initial phenotype screening and characterisation of the population, with respect to the whole data set and per gender. Figure 4.1 displays the distributions resulting from this analysis regarding patient's age, number of chronic diseases and number of consults attended.

With respect to the age spectrum of Dementia patients, presented in Figure 4.1 (a), it is possible to see that the distribution is very similar for the whole population, as well as for males and females individually. These patients oscillate in age between 58 and 102 years old, specifically between 61 and 102 for male patients and 60 and 100 for females. The first age quartile for all Dementia patients is set to 76 years old, meaning that the majority of Dementia patients (75%) are older than that. It is also possible to observe some points outside of the minimum age value defined by the box plot (58 years old), which represent patients identified as *outliers*, meaning that these patients fall out of the most common age distribution. Comparing female and male patients, it is possible to identify the existence of a younger male population with Dementia.

Moreover, in order to have a clearer idea of the age spectrum of Dementia patients, a bar plot with the patients' age distribution was obtained, which can be seen in Figure 4.2. It is clear that the majority of them have between seventy and ninety-five years old, as indicated by the box plots presented in Figure 4.1 and there are very few patients below the age of sixty. The age distribution of Dementia patients approximates a normal distribution, except for the fact that the mean age value is not located quite at the center of the distribution, but slightly tended to the left side, due to the existence of a few patients between the ages of twenty and sixty.

Moving on to the distribution of the number of chronic diseases that these Dementia patients suffer from, shown in Figure 4.1 (b), it is possible to see that the distribution is identical for the whole population



(a) Age distribution.

(b) Number of chronic diseases distribution.

(c) Total number of consults per patient.

(d) Number of different medical speciality consults per patient.

Figure 4.1: Distributions of phenotypes that characterize Dementia patients, for the whole population and by gender.

and for females, whilst for male patients there are slight differences. Dementia patients, in general, suffer between two and thirteen chronic diseases, as indicated by the minimum and maximum values of the box plot, being that male patients normally have at most ten. The median number of chronic diseases for this population is set to four, meaning that fifty percent suffer between two and four, while the remaining 50%, when considering female patients, suffer between four and thirteen chronic illnesses and males experience between four and ten. Male patients in this cohort that find themselves displaced from this distribution can have up to twenty-three chronic diseases, while females can admit up to eighteen.

Finally, regarding the distributions of the number of consults attended by Dementia patients, two distinct analysis were carried out. The first one considers all hospital visitations (Figure 4.1 (c)), while the second one only takes into consideration the number of different medical speciality consults attended by these patients (Figure 4.1 (d)). Focusing on Figure 4.1 (c), when considering all patients with Dementia, it is possible to observe that these visit the hospital for a medical consult between one and sixteen times, excluding *outliers*. The same behavior is observed for male patients, while female patients seem to have a slightly lower consult attendance, since the upper limit is set to thirteen consults. However, from the median values, it is possible to see that 50% of both female and male patients visit the hospital for a consultation a maximum of three times. Once again, it is possible to observe a considerable amount of

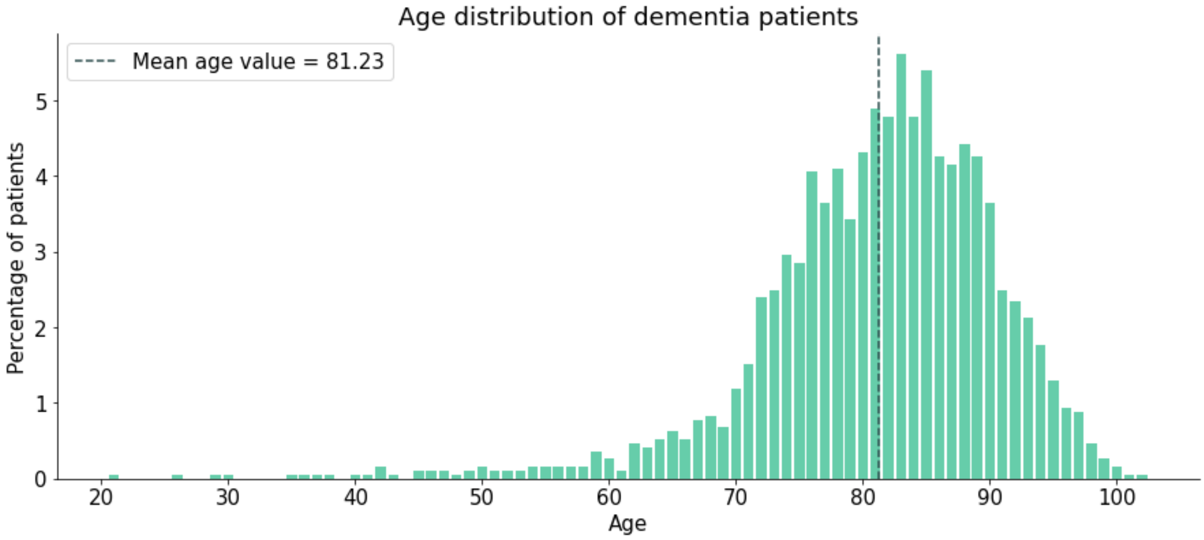


Figure 4.2: Age distribution of the population with Dementia.

patients identified as *outliers*, which had up to eighty-six consult attendances throughout the considered time period, when considering female patients and sixty when considering males. Finally, Figure 4.1 (d) shows that 50% of Dementia patients, considered as a whole or by gender, have at most two different medical speciality consults in their history. Furthermore, while for female patients the maximum number of different consults is set to six, for male patients this value is set to eight, which goes accordingly to the analysis done for Figure 4.1 (c) where it was observed that male patients tend to have more hospital visitations in their medical history. In the same line of thought, considering the quantiles defined by the box plots, 75% of female patients attended at most three different medical speciality consults, while male patients attended at most four different ones. Regarding the ones that fall out of this distribution, some male patients presented up to fifteen different medical speciality consults in their history, while females had up to fourteen.

Considering that Dementia was identified as presenting a certain heterogeneity in individual onset ages [24], a distribution of these patients age when diagnosed was obtained and can be seen in Figure 4.3. This distribution is very similar to the patient age distribution presented in Figure 4.2, being slightly shifted to the left hand side of the horizontal axis. Usually, Dementia is a disease which is diagnosed in patients with sixty-five years old or more, being that bellow this age it is considered a premature diagnosis. Cases of patients with onset ages between thirty and sixty-five years old are rare but possible. This early onset Dementia can be a result of post-traumatic experiences, substance abuse issues or genetic reasons. The U.S. Department of Health and Human Services published an article in the National Institute on Ageing journal stating that young-onset Dementia can be a consequence of an inherited mutation in one of three genes [45]. These mutations cause abnormal protein production which leads to the early development of symptoms. Due to the fact that it is rare to have a Dementia diagnosis in this age group, physicians do not always look for a Dementia diagnosis in such a young age, which may interfere with the process of early identification of this illness. Furthermore, its symptoms may overlap with those of psychological illnesses, such as depression, which once again may cause misdiagnosis or delay in the diagnosis of Dementia. Cases of onset in people with less than thirty years-old are very rare and have very few mentions in the literature. In our study cohort, three people with Dementia were diagnosed with less than thirty years old, one at nineteen, another at twenty-five and one with twenty-eight.

Since focusing on patients with multimorbidity, it was important to identify the chronic diseases present in this data set, as well as their prevalence and co-occurrence. There were one hundred and three distinct chronic illnesses identified amongst patients with Dementia, making it important to analyse which are the most prevailing ones, aiming at understanding which diseases may be more or less related to Dementia. Figure 4.4 represents the percentage of patients, by gender, that suffer from the ten most incident chronic diseases in the Dementia population, while Figures 4.6 and 4.7 show the graphs obtained to analyse the chronic diseases that this data set involves, as well as their co-occurrence. However, for relevance purposes, all these graphs represent the prevalence of diseases in the Dementia data set, considering only the co-occurrences that are present in more than two percent of the population.

Since there were many chronic diseases identified within this subset of patients, some of them having very low frequency, it was important to identify the most prevailing ones, aiming at understanding patterns regarding these patients and which diseases they suffer from. As it is possible to pinpoint through Figure 4.4, Hypertension (HTN), Dyslipidemia, Cerebrovascular disease, Obesity, Heart Failure, Chronic Kidney Disease (CKD), Atrial Fibrillation (AFib), Depression, Lumbago, Osteoarthritis, Thyroid disorders, Ischemic Cardiomyopathy (CM), Benign Prostatic Hyperplasia (BPH), Type 2 diabetes (T2DM) and Parkinson's disease represent the top fifteen incident illnesses in the Dementia population,

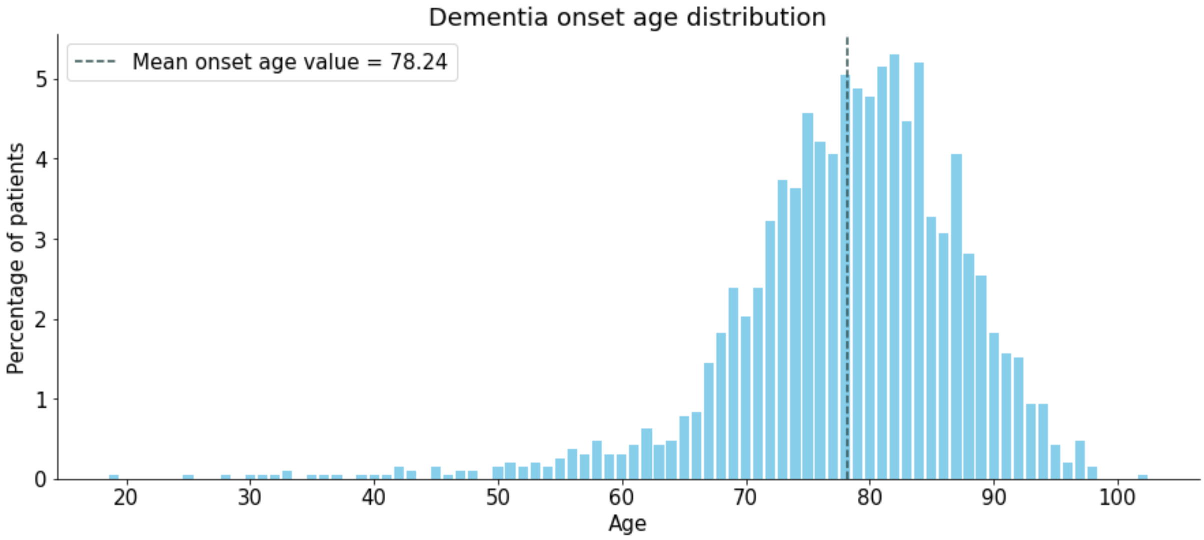


Figure 4.3: Dementia onset age distribution in the cohort under study.

presented by gender. It is possible to detect that certain diseases have more incidence in a certain gender group. For instance, cerebrovascular disease, heart failure, CKD, AFib, Ischemic CM, T2DM, Parkinson's and, of course, BPH have higher incidence in male patients when compared to female ones, while the prevalence of obesity, depression, lumbago, osteoarthritis and thyroid disorders is higher in females. The remaining diseases, HTN and dyslipidemia, are practically equally common in both gender groups. It is clear that vascular diseases have higher prevalence in males, while diseases related to pain (Ostheoartritis and Lumbgo) affect women more than men.

Considering what was mentioned in Section 2.3 regarding the comorbidities that have higher tendency to co-exist with Dementia, it is interesting to interpret these results in that sense. HTN, dyslipidemia, cerebrovascular disease, obesity, heart failure, AFib, Ischemic CM and T2DM are all comorbidities that represent risk factors for Vascular Dementia, since they represent diseases that affect blood vessels, leading to a poor brain irrigation, which leads to Dementia states. On the other hand, osteoarthritis and lumbago are illnesses that generate pain, which may cause difficulty to concentrate and to perform cognitively, leading a patient to loose certain basic functioning capabilities, hence, influencing Dementia states. Regarding depression, it may be dubious, since a depression diagnosis can be many times confused with an early Dementia state that goes undetected. This happens due to the fact that many Dementia states begin with behavioral changes, where patients feel and look debilitated, very much alike depression states, however, it may be an early signaling for Dementia. Finally, the fact that CKD, thyroid diseases and BPH belong to the top fifteen diseases suffered by Dementia patients is most probably not related with this disease, since the age group to which these patients belong to has an overall incidence of these diseases.

Furthermore, since some of these illnesses are very common within the elder population and in order to understand whether Dementia patients are more or less susceptible to suffering from these diseases

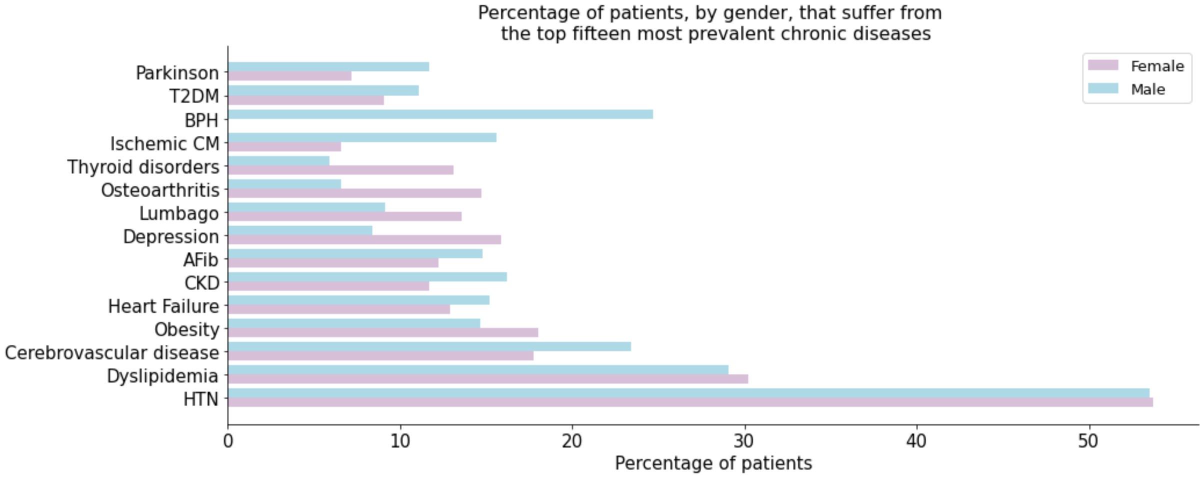


Figure 4.4: Top fifteen most prevalent chronic diseases within Dementia patients.

when compared to patients that suffer from MM in general, the ratio between the fraction of Dementia and multimorbidity patients that suffer from each one of the fifteen mentioned chronic diseases was obtained. Figure 4.5 represents the ratio between the fraction of Dementia and MM patients in general, that suffer from each one of the top fifteen chronic diseases, which tells us how much more likely a patient with Dementia is to develop a certain disease when compared to the overall MM population. A ratio around one says that the prevalence of that disease is practically the same whether a patient suffers from Dementia or not, hence, the origin of the plot was set to unity. A ratio below one indicates that Dementia patients are less susceptible of developing that co-morbidity when compared to MM patients. On the other hand, a ratio higher than one means higher odds of that disease co-existing with Dementia. For instance, the highest ratio is relative to Parkinson's disease, which says that a patient with Dementia is eight times more likely of also suffering from Parkinson than a patient with MM but that does not suffer from Dementia. This is in agreement to what was mentioned in Section 2.4, which is that it has been proven that patients with Parkinson's disease are much more susceptible to developing Dementia symptoms.

Moreover, despite the fact that certain risk factors, such as obesity, can be associated to cognitive decline, this analysis demonstrates that Dementia patients, at least the ones that are part of this cohort, are less associated to this risk factor than the MM population, since a ratio below unity is obtained. The same is observed for lumbago and thyroid diseases. Despite belonging to the top fifteen most incident chronic conditions in the Dementia population, their prevalence is not necessarily associated with this disease. The fact that lumbago, which is described as acute lower back pain, and thyroid disorders

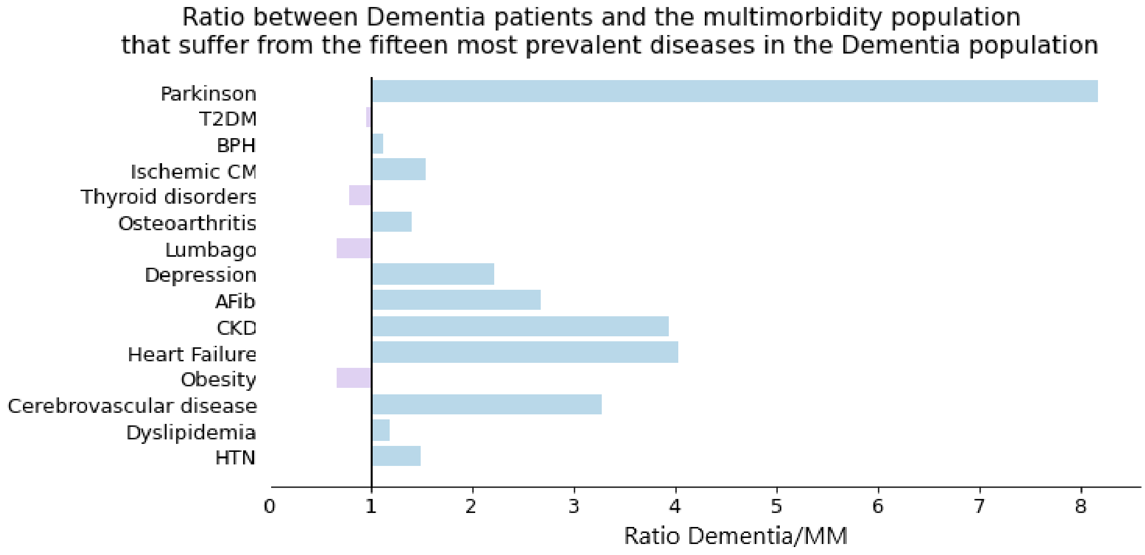


Figure 4.5: Incidence of the top fifteen chronic diseases in Dementia patients relative to the MM population. The vertical line represents the unity threshold from which Dementia patients are more susceptible than MM patients of suffering from one of the top fifteen diseases.

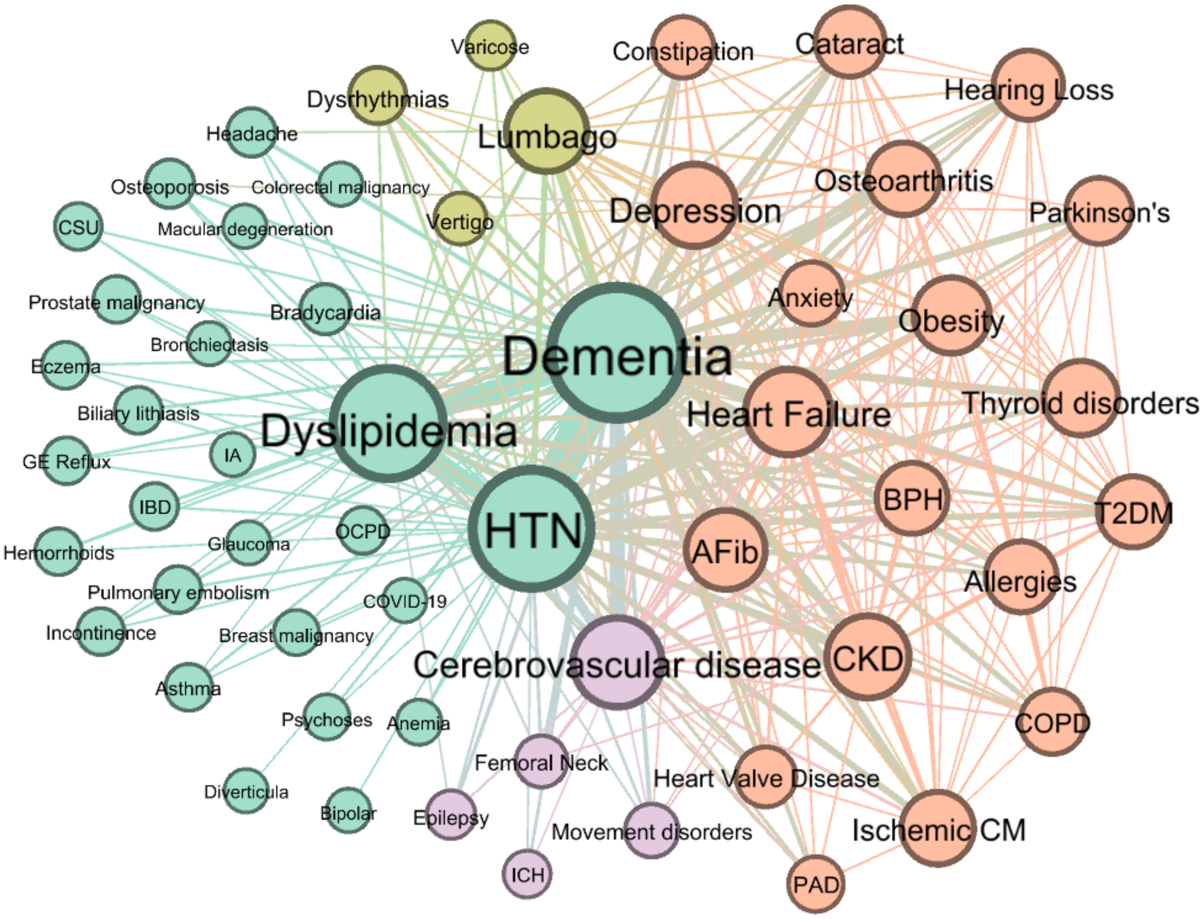


Figure 4.6: Graph of chronic disease co-occurrence in Dementia patients, considering only the pair of illnesses that have incidence in more than one percent of the population. Bigger nodes indicate higher incidence of the chronic disease in the data set and wider edges indicate more co-occurrence of that pair of diseases.

represent two chronic diseases that have high incidence in the elder population [46] [47] can justify the fact that these belong to the top fifteen common diseases in Dementia patients but the probability of these diseases co-exiting with Dementia is not higher than the probability of MM patients suffering from it.

Additionally, as previously mentioned, there is a type of Dementia, which is the vascular one, where patients that suffer from a vascular disease have higher probability of developing Dementia symptoms. It is possible to verify this preface by looking at the ratios corresponding to cerebrovascular disease, heart failure, CKD and AFib, which represent risk factor for Vascular Dementia, since these patients are two to four times more probable of suffering from Dementia and one more of these illnesses, when compared to the general MM population. Finally, depression has also been previously linked to Dementia and with this analysis it is possible to see that it is 2.22 times more probable that depression co-exists with Dementia than with other disorders in general.

Shifting the focus to the chronic disease co-occurrence graphs obtained for the whole population (Figure 4.6) and by gender (Figure 4.7), it is important to point out that bigger nodes indicate higher prevalence of that chronic illness within Dementia patients, while wider edges indicate more co-occurrence of a pair of diseases. The co-occurrence graph obtained for all Dementia patients was filtered at one percent, meaning that what is observable in Figure 4.6 is only relative to diseases that co-occur in more than one percent of the population. On the other hand, graphs obtained by gender were filtered by two percent, in order to perform a more specified analysis. It is clear from the edges connecting Dementia to other chronic illnesses in the three graphs, that the ones more often in co-occurrence with Dementia are HTN and dyslipidemia.

Regarding Figure 4.6, it is possible to see the existence of four distinct disease sub-groups. These were obtained by resorting to the modularity property, which is a measure of a network's or graph's structure, assessing the degree to which these can be divided into separate communities that have higher interaction between them when compared to others [48]. It is interesting to see that the majority of the top fifteen previously identified chronic diseases are highly interconnected in the orange sub-group. The top two chronic conditions, besides Dementia, which are HTN and dyslipidemia belong to the same sub-group, together with the remaining diseases. In addition, it is possible to observe that cerebrovascular disease and lumbago, despite being in the top fifteen comorbidities of Dementia patients, do not belong to to the same community as the remaining ones. Lumbago was identified as being more interconnected with vertigo, varicose and dysrhythmias, while cerebrovascular disease is more related to *Ichthyophthirius multifiliis* (ICH), femoral neck pathologies, epilepsy and movement disorders.

Focusing on the co-occurrence graphs by gender, it is curious to observe the differences in chronic disease co-occurrence. For instance, it is possible to verify by comparing these graphs that diseases related to the vascular system are more common in men, while women are more often affected by depression, diseases related to pain, such as lumbago and osteoarthritis, and thyroid disorders. Looking at Figure 4.7 and comparing the position of the nodes corresponding to CKD, Ischemic CM and AFib in the male and female graphs, it is possible to verify higher incidence and co-occurrence of these diseases with Dementia in male patients. On the other hand, depression in female patients is identified as the fourth most common comorbidity, while for male patients this disease belongs to the second half of most prevalent illnesses. The same goes for the pain conditions, where we can see that osteoarthritis and lumbago appear as two of the most incident comorbidities in the female population, while for males these are two of the least existing co-occurring diseases.

To finalize this initial phenotype screening process, the prevalence of each medical speciality present in the data set was assessed. A population pyramid with this information is presented in Figure 4.8. It is interesting to observe that the prevalence of each of the medical speciality consults is similar for

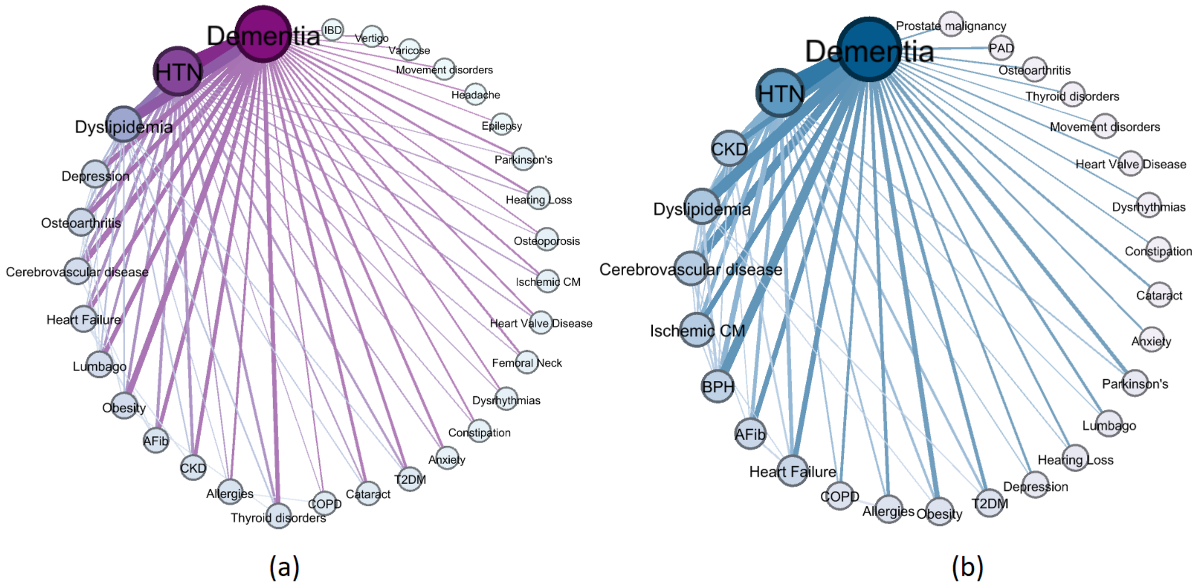


Figure 4.7: Graph of chronic disease co-occurrence in (a) female and (b) male Dementia patients, considering only the pair of illnesses that have incidence in more than two percent of the population. Bigger nodes indicate higher incidence of the chronic disease in the data set and wider edges indicate more co-occurrence of that pair of diseases. Nodes are ordered in a circular way by the most prevalent chronic disease to the least prevalent.

both female and male patients. Moreover, being a neurological disease, it is immediate that Neurology consults are the most occurring consults amongst Dementia patients. As it is possible to observe, a great part of the population (approximately 72%) had at least one Neurology consult throughout their hospital activity. Beyond Neurology consults, General and Family Medicine (GFM), Internal Medicine, Anesthesiology and Cardiology consults are very present in these patients history, while Immunoallergology, Hematology, Rheumatology, Nephrology and Dental are the five less attended medical speciality consults.

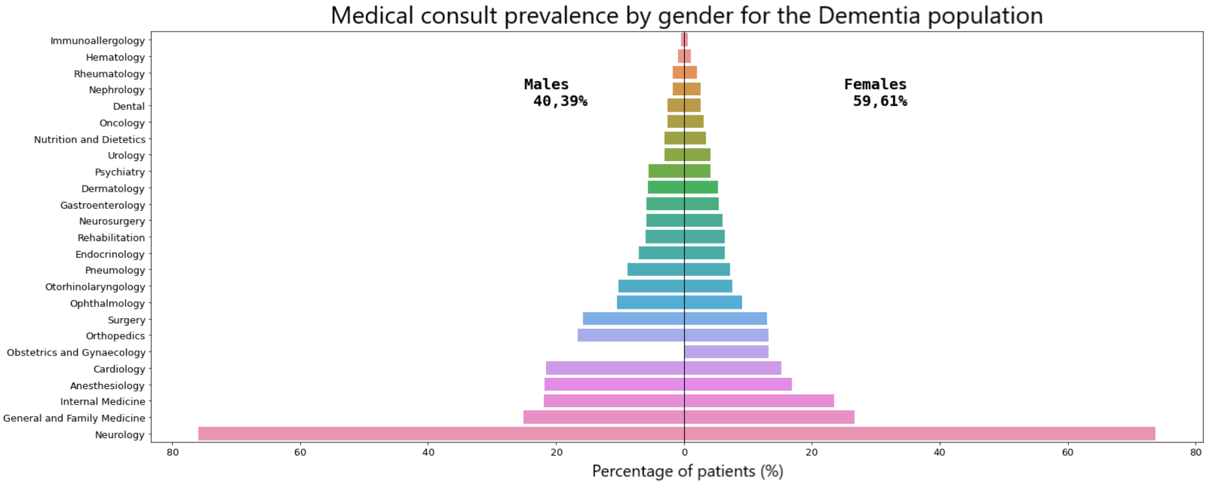


Figure 4.8: Percentage of patients, by gender, that attended each medical speciality consult at least once.

4.2 Clinical pathway analysis

Shifting the focus from the phenotype evaluation of the population with Dementia to their hospital activity regarding medical consult attendance, a Markov chain and a patient stratification approach were used in order to detect prevailing patterns within these patients.

First of all, it is important to point out the fact that within the 1 924 patients with Dementia present in the study cohort, 59 of them did not have information on their medical consults. For this reason, it was not possible to include these patients in this part of the study, having remained 1 865 Dementia patients for the clinical pathway analysis carried out.

4.2.1 Markov Chains

In this first stage, two transition matrices were initially obtained and visualized in heatmaps, as shown in Figure 4.9, for the two mentioned cases:

- Considering consecutive transitions between the same medical speciality consult, as presented in Figure 4.9 (a), for which only patients with one consult in their history were filtered, remaining then 1 607 Dementia patients, from which 951 are female and 656 male, from the 1 865 patients considered for this analysis.

- Not considering consecutive transitions between the same medical speciality, given by Figure 4.9 (b), aggregating consecutive occurrences of the same consult into one and filtering patients who ended up with one consult alone, remaining 1 204 patients, 709 females and 495 males, to be considered for the transition matrix formulation.

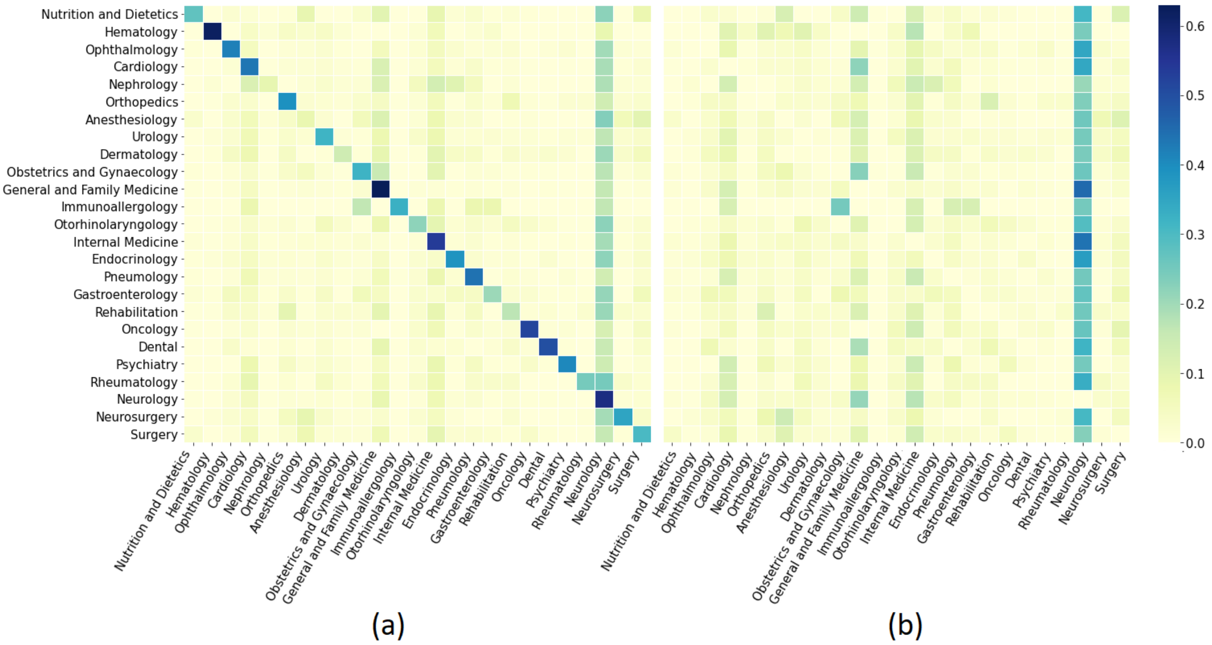


Figure 4.9: Heatmaps displaying the transition probabilities of Dementia patients moving between medical speciality consults, when (a) considering and (b) not considering consecutive transitions between the same medical speciality consult.

It is important to keep in mind, when analyzing these heatmaps, that each entrance of a transition matrix is the probability of moving to the column event, knowing that the previous event is represented by the row of the matrix. Also, a transition matrix representing a Markov chain is a stochastic matrix, meaning that each row represents the most probable transitions when the origin is the consult indicated by the horizontal entrance.

With respect to the first scenario, shown in Fig. 4.9 (a), where consecutive transitions between the same medical speciality are considered, it is clear that the most prevailing transitions are between the same consult. This may be indicative of the fact that patients generally have a follow-up consult in the same medical speciality prior to being redirected to a different one. It is also interesting to notice that, since dealing with a neurological illness, independently of the source consult, one of the most probable transitions is to a Neurology appointment. Furthermore, despite not being as clear, it is also possible to observe a slight tendency of Dementia patients of transitioning to Internal Medicine and GFM consults, which goes accordingly to the fact that both these medical specialities are prevailing amongst Dementia patients.

Regarding the second scenario in Figure 4.9 (b), where consecutive transitions between the same medical consult are not considered, it is still clear that transitions to Neurology appointments still prevail, no matter what the consult of origin is. Transitions to Internal Medicine, GFM and Cardiology consults can also be prevailing, which makes sense considering the prevalence of these medical specialities in

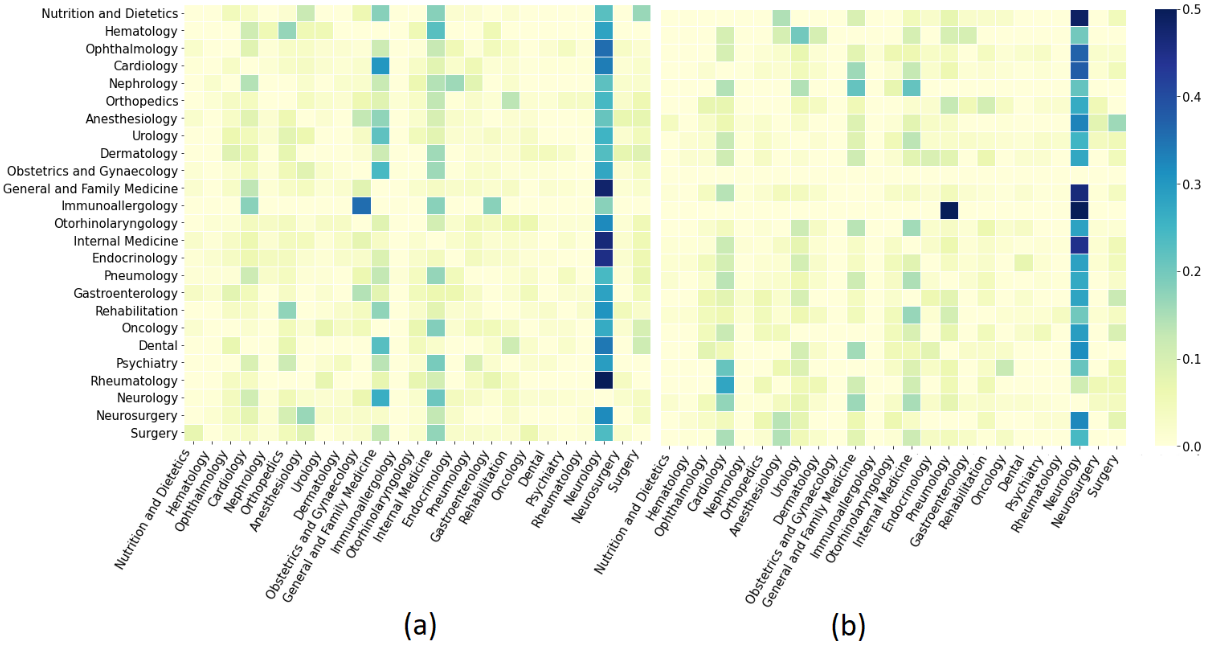


Figure 4.10: Heatmaps displaying the transition probabilities of (a) female and (b) male Dementia patients moving between medical speciality consults, not considering consecutive transitions between the same consult.

the data set.

Furthermore, the same analysis was carried out, but by gender. When considering consecutive visits to the same medical speciality consultation, the same patterns were observed as for the whole population. This is, a clear prevalence of transitions between the same medical speciality is detected. On the other hand, regarding the scenario where these consecutive transitions are not taken into consideration, there are slight differences in the patterns observed by gender, as represented in Figure 4.10. Patterns with respect to transitions to Neurology consults are still observable in both gender groups. Now, looking at Figure 4.10 (a), representing the heatmap for the female patients, it is possible to see stronger patterns of redirection to GFM and Internal Medicine, when compared to male patients (Figure 4.10 (b)). On the other hand, male patients tend to undergo more transitions to Cardiology and Urology consults, when compared to female patients. On the other hand, a slight increase in transitions to Orthopedics consults is observable for female patients, when compared to male.

4.2.2 AliClu

Still aiming at identifying consult activity patterns within Dementia patients, the AliClu algorithm was used in order to stratify these patients. This section presents the results of this clustering process, as well as of the whole pipeline regarding pre-processing the data and optimizing the algorithm parameters.

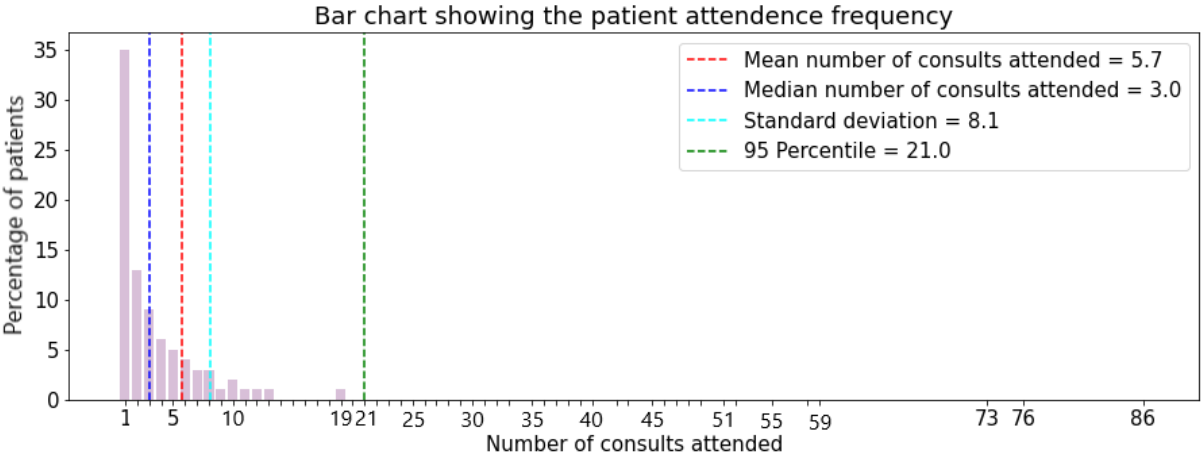


Figure 4.11: Overview of the number of consults attended by the 1 865 patients suffering from Dementia.

4.2.2.A Pre-processing

Prior to using the AliClu algorithm, it was necessary to take the raw data handed in panel data format in order to obtain a temporal sequence for each patient. Hence, the pre-processing script available alongside AliClu was applied in the available data, which resulted in a two-column file with the information on the patients ID, as well as their respective temporal sequence regarding medical consults attended in HLL. The last step of the pre-processing stage is filtering some of these patients that do not fit the most appropriate criteria to undergo the clustering algorithm.

As previously mentioned, the focus is on patients with multimorbidity, hence, it is only interesting to analyse the activity of these patients regarding two or more medical specialities. Figure 4.11 shows the distribution of the number of consults attended by the patients in the data set being analysed. It is important to point out that this bar chart serves the purpose of understanding the amount of hospital visits regarding these patients, hence, the number of consults displayed includes the re-occurrence of certain medical speciality consults, if it's the case. However, if the re-occurrence is in a row, it will only be taken into consideration as one consult, since AliClu considers the same event in a row as one.

As it is possible to observe, approximately 660 patients, roughly 35%, have at most one medical speciality consult in their history. The red vertical line represents the mean number of patient hospital visitations, from which it is possible to infer that a patient with Dementia, in this specific data set, goes to an average of five to six consults during a period as long as the one being considered. On the other hand, three is the median number of consults attended by these patients, represented by the dark blue line, meaning that 50% of this population visited the hospital for an appointment at most three times, while the remaining 50% visited between three and eighty-six times. Furthermore, the green vertical line, represents the 95 percentile, which is twenty-one, indicating that 95% of the population visited the hospital for a medical consult twenty-one times or less.

Table 4.1: List of parameters analysed, using the modified version of AliClu.

Gap penalties, *g*	-0.5, -0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4, 0.5
Temporal penalties, *Tp*	1, 2, 5, 7, 10
Number of clusters, *k*	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

In view of what was mentioned in the previous section, after obtaining the PE sequence for each of the 1 865 patients, it was necessary to filter these patients before moving on to the clustering process. Thus, patients with only one medical consult and patients who fall into the 5% over the 95 percentile, are filtered from the data set, remaining 1 118 patients with Dementia available for undergoing the sequence alignment and clustering algorithm.

4.2.2.B Parameter optimization

So as to understand which AliClu parameters would give the best patient clusters, a range of gap penalties (*g*), temporal penalties (*Tp*) and number of clusters (*k*), presented in Table 4.1 were analysed, resorting to the modified version of the AliClu algorithm, which returns the average SS for the set of clusters obtained with each combination of these three parameters. This analysis began by randomly partitioning the filtered data set, which includes 1 118 patients with Dementia, into six equal-sized subsets. Each subset of 186 patients was then subject to this optimum parameter search. Hence, for each partition, the average SS for each combination of AliClu parameters was obtained.

The results of the optimum parameter search are graphically presented as the average SS as a function of the growing number of clusters formed. A figure is generated for each gap penalty evaluated, containing a line plot for each temporal penalty. Each point in the plot corresponds to the mean average SS across the six considered partitions. In Figure 4.12 it is possible to see these results for gap penalties of 0.1 and 0.2, which were the ones that presented the best outcomes. The plots obtained for the remaining gap penalties considered in the parameter optimization process are available in Appendix A. Negative gap penalties resulted in very low values of average SS. In addition, visual inspection of the clusters obtained using these gap penalties lead to the conclusion that these were not the most appropriate values for this parameter. Regarding the results for positive g, it became clear that gap penalties of 0.1 and 0.2 performed better both in terms of average SS, as well as by observation of the clusters.

Regarding the influence of the temporal penalty in the results, there is a clear increase in the values of average SS proportional to an increase in *Tp*. Comparing both images in Figure 4.12, it is possible to see that, in a general way, using *g* = 0.1 leads to higher average SSs, reaching values between 0.25 and 0.3 for the three higher *Tp* values analysed. However, for a gap penalty of 0.2 good results are also reached. Furthermore, it is also interesting to observe how the average SS evolves with the growing

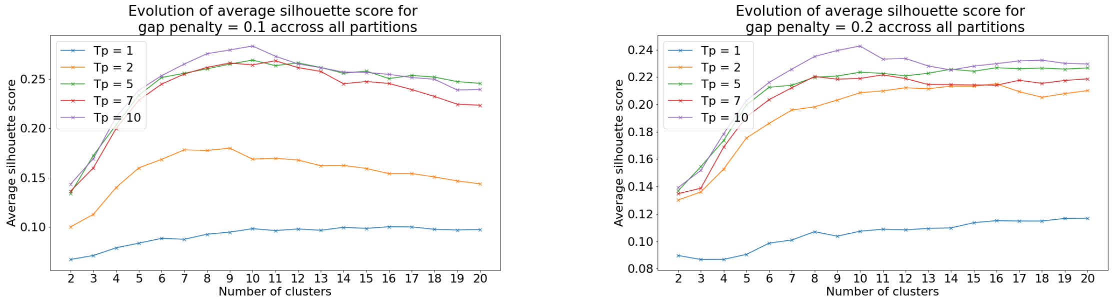
Table 4.2: Summary table of the optimum parameters per gap penalty and corresponding average SS.

Gap penalty, *g*	Temporal penalty, *Tp*	Number of clusters, *k*	Average Silhouette Score
-0.5	7	20	0.137433
-0.4	10	2	0.136294
-0.3	10	2	0.0981721
-0.2	2	2	0.055397
-0.1	10	8	0.0932941
0.1	**10**	**10**	**0.283154**
0.2	**10**	**10**	**0.242617**
0.3	10	20	0.213723
0.4	2	19	0.195603
0.5	1	2	0.235053

number of clusters. For a gap penalty of 0.1, there is an initial increase, reaching the maximum average SS around *k = 11* and finally a slight decrease is observed. On the other hand, for a gap penalty of 0.2, there is an initial more abrupt increase in the average SS until *k = 11*, from where a stabilization of the average SS is detected until *k = 20*.

Table 4.2 summarizes which parameters (*g*, *Tp*, *k*) lead to the best SS, per gap penalty. As previously observed from the obtained plots, a higher temporal penalty and gap penalties of 0.1 and 0.2 lead to the best results. Regarding the optimum number of clusters, it can be quite heterogeneous, as expected, since besides being data dependent, the choice of the remaining parameters can also influence. For instance, for gap penalties of -0.4, -0.3, -0.2 and 0.5, the optimum number of clusters is *k = 2*, however this choice of clusters is not reliable, considering the amount of data involved. Additionally, visual inspection of the clusters obtained lead to the conclusion that no matter the chosen parameters, when considering only two groups of patients, these would end up not being well distributed, since one of them would form one cluster, while the remaining ones would form the other.

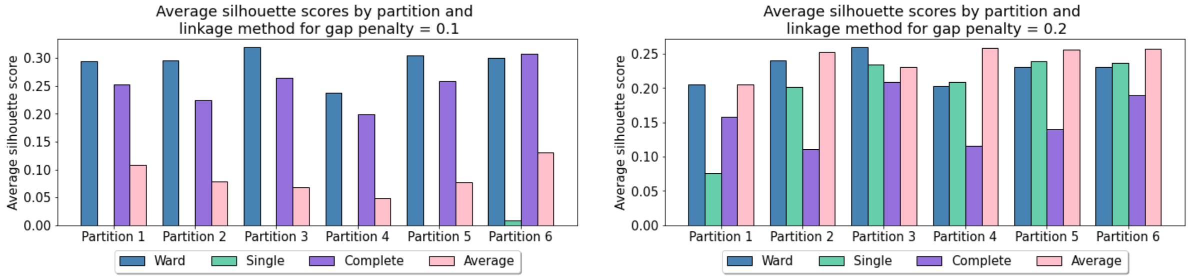
Besides evaluating the cluster quality by the average SSs obtained, a visual inspection was also



(a) Gap penalty = 0.1

(b) Gap penalty = 0.2

Figure 4.12: Evolution of the average SS for each combination of parameters (*g*, *Tp*, *k*), for gap values of 0.1 and 0.2 and all values of *Tp* and *k* presented in Table 4.1



(a) Gap penalty = 0.1

(b) Gap penalty = 0.2

Figure 4.13: Average SS values for the best parameters (*g*, *Tp*, *k*) obtained for each partition using the four different linkage functions available.

performed to guarantee the groups were logically formed. This analysis supported the SS results, in a way that the most robust clusters were the ones achieved for lower positive gap penalties (0.1 and 0.2) and a higher temporal penalty of 10. Regarding the optimum number of clusters for the whole data set, this choice is crucial and should be chosen depending on the subset aiming at stratifying. That being said, this parameter was chosen ahead, when applying the AliClu algorithm to the whole set of patients to obtain the desired final stratification.

Having undergone this parameter optimization process for the three most crucial AliClu parameters, next steps involved confirming the initial hypothesis regarding secondary parameters of the algorithm. For this purpose, the parameters already explored were set to the values which resulted in the top two values of average SS, that is, gap penalties of both 0.1 and 0.2 were considered, while the temporal penalty and number of clusters were set to 10.

Firstly, in order to corroborate the hypothesis that Ward's linkage function is indeed the best option for hierarchical clustering, AliClu was tested once again for the six partitions, this time varying the linkage method used, between Ward, single, complete and average link functions. The results are presented in the bar charts given by Figure 4.13. Note that the bars corresponding to the single linkage function don't always appear for a gap penalty of 0.1, due to the fact that the resulting average SSs are negative. Paying attention to Figure 4.13 (a), it is possible to see that the complete linkage function performed slightly better than the Ward's method for partition number 6 and (*g* = 0.1, *Tp* = 10, *k* = 10). However, for the remaining partitions, the method that proved better for the hierarchical clustering was, as expected, the Ward's method. For a gap penalty of 0.2, the average link function performed slightly better than Ward's for almost every considered partition, indicating that this linkage method is also a good choice, at least when considering *g = 0.2*. However, better results based on average SS were reached for *g = 0.1* using Ward's method.

Moreover, it was also hypothesized that gaps ranging from -1 to -0.6 and 0.6 to 1 were not as appropriate for the sequence alignment process as gaps ranging from -0.5 to 0.5, due to a higher possibility of obtaining total misalignments. To confirm this assumption, AliClu was once again run with these gap values, a temporal penalty of 10 and a total number of 10 clusters, which were the parameters that proved to lead to best results. Figure 4.14 shows the average SS obtained across all partitions.

Contrarily to what was observed when testing gaps between -0.5 and 0.5, in the case of gaps represented in Figure 4.14, the negative ones present higher average SSs than positives. Nevertheless, comparing these values with the ones previously obtained, it is guaranteed that these gap penalties do not lead to better results. As it is possible to see, the maximum average SSs obtained with this analysis do not exceed 0.14, while for gap penalties previously tested, these scores reach values up to 0.3. Hence, the assumption that these gap penalties lead to poorer clusters is confirmed. In addition, to further assure that these formed clusters were not eligible, a more careful verification was pursued, which lead to the conclusion that the clusters themselves formed with these gap penalty values were not satisfactory at all, as expected. For negative gap penalties between -1 and -0.6, all clusters end up with only one element, except for one of them, which ends up having practically all elements. This phenomenon justifies the fact that, for negative gap values presented in Figure 4.14, the average SSs across all partitions are very close, since in those situations, practically all elements end up in only one cluster. A similar pattern is verified for gap penalties between 0.6 and 1, where the elements end up not being well distributed, however, not as when these gaps are negative.

Finally, the last hypothesis to be verified was regarding the number of bootstrap samples for the clustering part of the algorithm. The base value set for this parameter was chosen according to [39] and set to *M = 250*. In order to guarantee that a higher number of bootstrap samples would not lead to better results, once again the algorithm was applied, this time choosing the best values of *k*, *Tp* and *g*, varying the number of bootstrap samples, as shown in Figure 4.15. It is possible to see that, independent of the value given to *M*, the average SS for all partitions is constant. This leads to the conclusion that assigning a higher number of bootstrap samples to the algorithm is not worth it in terms of the quality of

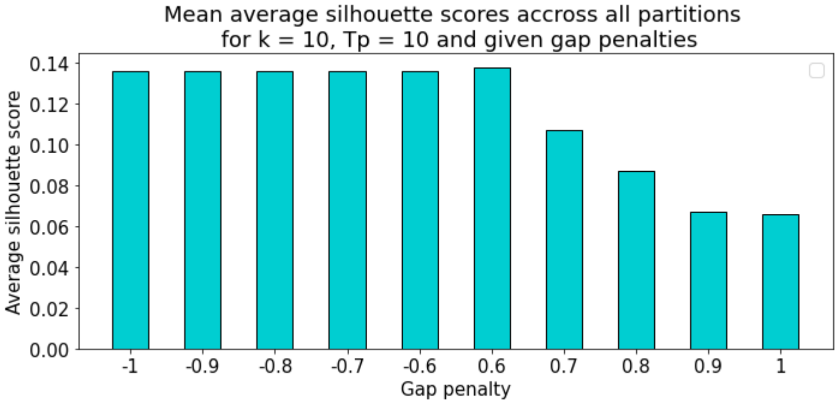
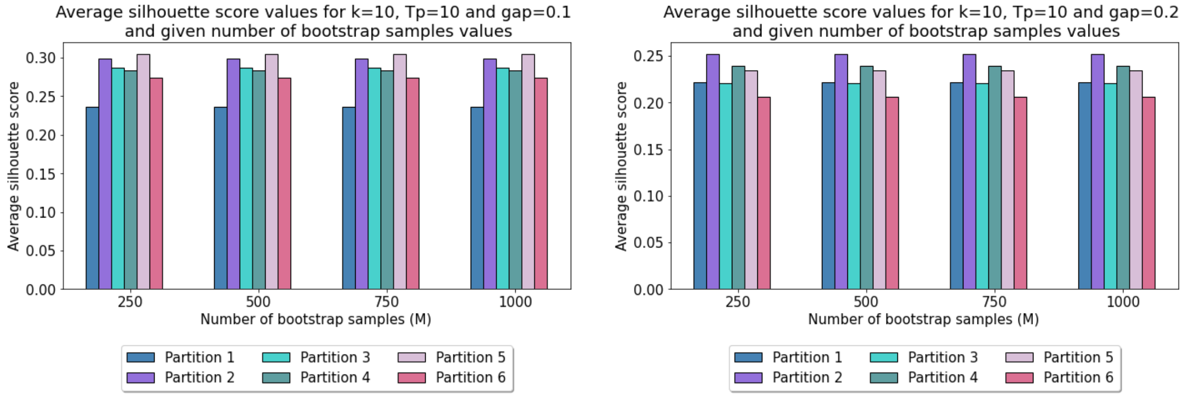


Figure 4.14: Average SS across six Dementia data set partitions, for when the temporal penalty and number of clusters are set to 10 and gap penalties range from -1 to -0.6 and 0.6 to 1.



(a) Gap penalty = 0.1

(b) Gap penalty = 0.2

Figure 4.15: Average SS values for the top two values for the gap penalty, *Tp = 10* and *k = 10*, obtained for each partition using various number of bootstrap samples, *M*.

the clusters formed, only making the algorithm computationally heavier, increasing its running time.

4.2.2.C Obtaining the final clusters

Subsequently to reaching a conclusion about which parameters would provide the best clusters when applied to the whole dataset, once again, the AliClu algorithm was ran with all 1 118 patients. However, due to the fact that the optimum parameters are in a great way influenced by the cohort itself and more

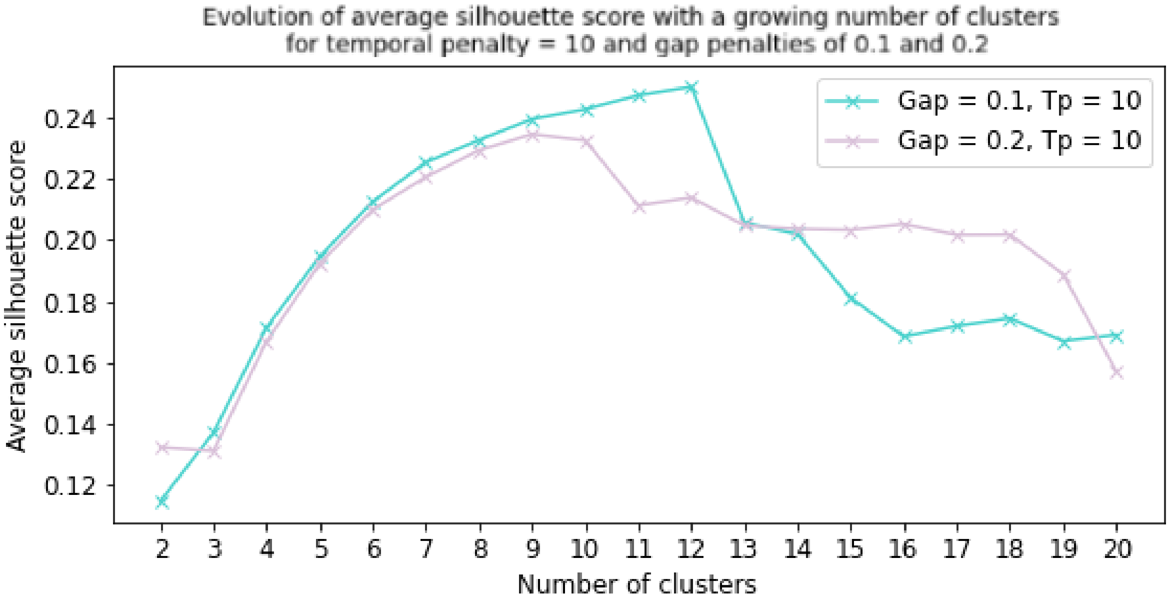


Figure 4.16: Evolution of the average SS with the number of clusters formed from *k = 2* to *k = 20*, for a *Tp* of 10 and gap penalties 0.1 and 0.2

than one value for some AliClu parameters performed well in the parameter optimization process, two values for the gap penalty and a range of *k* from 2 to 20 were chosen to run AliClu with the whole data set. The temporal penalty was set to 10, while the gap penalties chosen were 0.1 and 0.2. The Ward's linkage method for the hierarchical clustering was used and the number of bootstrap samples was set to 250.

Once again, an adaptation of the original AliClu was used in order to allow interpretation of the SSs for each combination of parameters tested. Figure 4.16 shows the change in SS with the growing number of clusters formed, considering the mentioned parameters.These plots show that the maximum average SS is reached when *g* is set to 0.1, *Tp* to 10 and *k* to 12 clusters, yielding a value of approximately 0.25. Thus, the 12 clusters obtained with these established parameters were chosen to be further analysed.

Still in the context of confirming the parameters choice and to make sure that other temporal penalty values, besides the ones tested in the parameter optimization process, would not yield better results, we did one last verification. Specifically, the gap penalty was set to 0.1 and the number of clusters to 12, while the temporal penalty was tested for a range of values from 1 to 15 with a step of 1. Figure 4.17 shows the evolution of the average SS considering these parameters. Interpretation of the plot shows a more steep increase in the SS until *Tp = 6*. From this point on, until the highest tested temporal penalty, the variation is not significant and the average SS stabilizes. Visual observation of the clusters leads to the conclusion that the clusters formed from the point where *Tp = 6* until *Tp = 15* are practically the same. This slight increase in the average SS with the increase in *Tp* may be justified by the fact that this metric depends on the distance matrix resultant from the alignment process, which in turn is proportional to the value attributed to the temporal penalty. Hence, the larger the temporal penalty, higher SSs will be registered, despite the differences observed in the groups obtained not being significant.

It was curious to notice that the algorithm grouped the patients mainly by the first consult registered in their medical history. Table 4.3 shows the medical speciality indicating the start of the patients' pathways within each cluster, which is typically the dominant one, as well as the number of elements that form each cluster. Apart from cluster 9, it is clearly noticeable that there is a medical speciality consult that dominates each of the clusters. This cluster is formed by *outliers*, namely, elements which do not fit any of the remaining ones or the elements that the algorithm was not able to properly align or find a proper alignment pair.

Despite the clear tendency of the algorithm to group patients mainly based on the initial consult, it is possible to observe other patterns throughout the sequences belonging to the same cluster, since during the pairwise sequence alignment process, the algorithm aligns the PE sequences, inserting gaps where they do not match, instead of inserting a mismatch.

Moreover, in order to quantitatively analyse the content present in each one of the twelve clusters formed with (*g* = 0.1, *Tp* = 10, *k* = 12), it is essential to look at Figure 4.18, where the SS for each sample

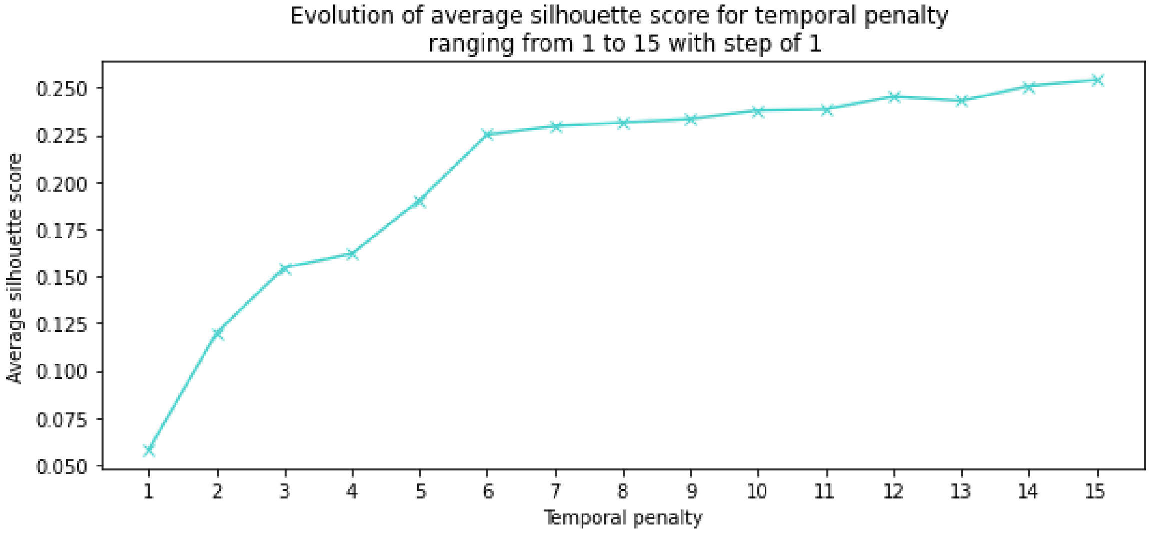


Figure 4.17: Evolution of the average SS for a gap penalty of 0.1, number of clusters of 11 and varying temporal penalty between 1 and 15 with a step of 1.

Table 4.3: Number of elements and prevailing medical speciality consult within each cluster.

Cluster label	Prevailing medical speciality consult	Number of elements
1	Endocrinology	17
2	Nutrition and Dietetics	23
3	Orthopedics	24
4	Pneumology	31
5	Obstetrics and Gynaecology	33
6	Surgery	62
7	Cardiology	78
8	Anesthesiology	104
9	*Outliers*	108
10	Internal Medicine	150
11	Neurology	227
12	General and Family Medicine	261

in the data set, as well as the average SSs for each cluster and across all samples is represented. Keeping in mind that the SS measures how close a sample is to elements belonging to the same cluster, in comparison with how close or far that same sample is from other elements in different clusters, it is possible to verify that cluster 9 is composed by *outliers*. This conclusion is reachable due to the fact that the SSs of all samples in this cluster have very low values, being that the majority of the elements even have negative scores. Furthermore, it is also possible to identify *outliers* in several other clusters, considering that some elements of clusters 2, 6, 8, 9, 10 and 12 possess negative SSs.

The vertical red dotted line represents the average SS across all samples of the data set and it is interesting to identify where each group stands in terms of its own average SS. Except for clusters 2, 4 and 9, all remaining ones have average SSs higher than the overall score. Cluster 9 was expected to be below average since it is composed by *outliers*, while 2 and 4 were not. However, despite being below average, the difference is not much and may be justified by the smaller amount of elements that compose them.

Furthermore, considering the values of the clustering indices which are already part of AliClu, more specifically Rand, Adjusted Rand, Fowlkes and Mallows, Jaccard and Adjusted Wallace, calculated for each set of clusters formed from k = 2 to k = 20, during AliClu's bootstrapping validation process, it is possible to see that the index values for the chosen optimum number of clusters (k = 12), presented in Table 4.4, are not maximum. The five indices are maximum for *k = 2*, however, patients end up not being well distributed, as previously verified. Patients who begin their hospital activity in GFM form one group, while the remaining ones form the other. It is curious to notice that until *k = 12*, the indices in question round the same values, reaching a maximum when *k = 11*, from where a decrease is noticed. Furthermore, through visual inspection of the clusters, it is possible to observe the consecutive grouping of patients which have the same first medical appointment. Specifically, as k increases, the algorithm successively groups a set of patients who present the first event in common. Curiously, from the moment k reaches 12, instead of creating a new cluster with a different opening consult, other previously formed groups for lower k values are split up. That being said and coupled with the fact that the index values for *k = 12* are very satisfactory, despite not being maximum, choosing this value as the optimum number of clusters was considered the most appropriate choice.

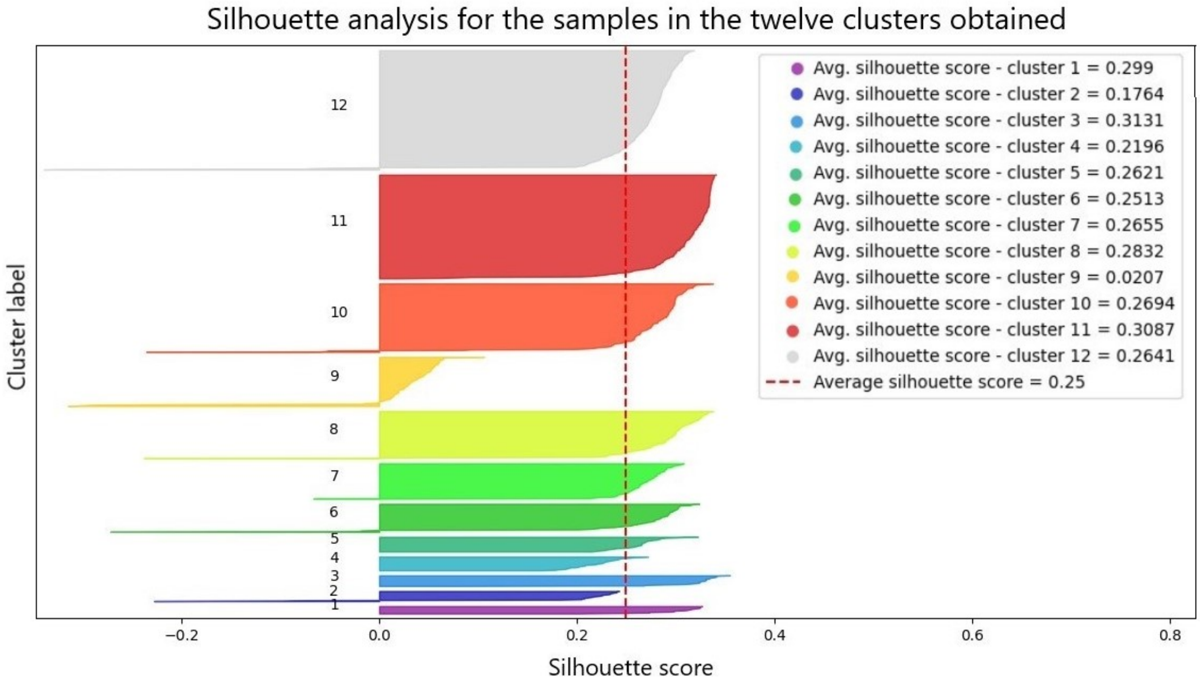


Figure 4.18: Silhouette scores for each sample present in each of the obtained twelve clusters for g = 0.1 and Tp = 10, average SS per cluster and across all samples in the data set.

In order to further validate the groups obtained, the quantitative cluster stability analysis incorporated in AliClu, which uses the same bootstrap method used for the choice of the best *k*, is analysed. As previously mentioned, for each of the formed clusters, the average, median and standard deviation of

Table 4.4: Values of Rand, Adjusted Rand, Fowlkes and Mallows, Jaccard and Adjusted Wallace clustering indices for each set of *k* clusters formed with (g = 0.1, Tp = 10).

Number of clusters, *k*	Rand	Adjusted Rand	Fowlkes and Mallows	Jaccard	Adjusted Wallace
2	0.99	0.977	0.992	0.984	0.978
3	0.984	0.968	0.981	0.963	0.963
4	0.984	0.961	0.973	0.947	0.957
5	0.983	0.953	0.964	0.931	0.951
6	0.986	0.956	0.965	0.932	0.956
7	0.987	0.954	0.962	0.928	0.957
8	0.98	0.928	0.94	0.888	0.928
9	0.988	0.955	0.962	0.927	0.957
10	0.985	0.943	0.951	0.908	0.943
11	0.987	0.946	0.954	0.912	0.944
12	0.98	0.916	0.928	0.866	0.896
13	0.976	0.892	0.906	0.829	0.883
14	0.972	0.862	0.879	0.784	0.899
15	0.973	0.864	0.879	0.786	0.888
16	0.974	0.866	0.881	0.787	0.892
17	0.976	0.871	0.885	0.794	0.887
18	0.978	0.878	0.89	0.803	0.883
19	0.976	0.865	0.878	0.784	0.884
20	0.976	0.863	0.876	0.781	0.871

Table 4.5: Cluster stability assessment through the median, average (avg) and standard deviation (std) of the Jaccard, Dice coefficient and recovery rate correspondence measures for clusters obtained with (k = 12, g = 0.1, Tp = 10).

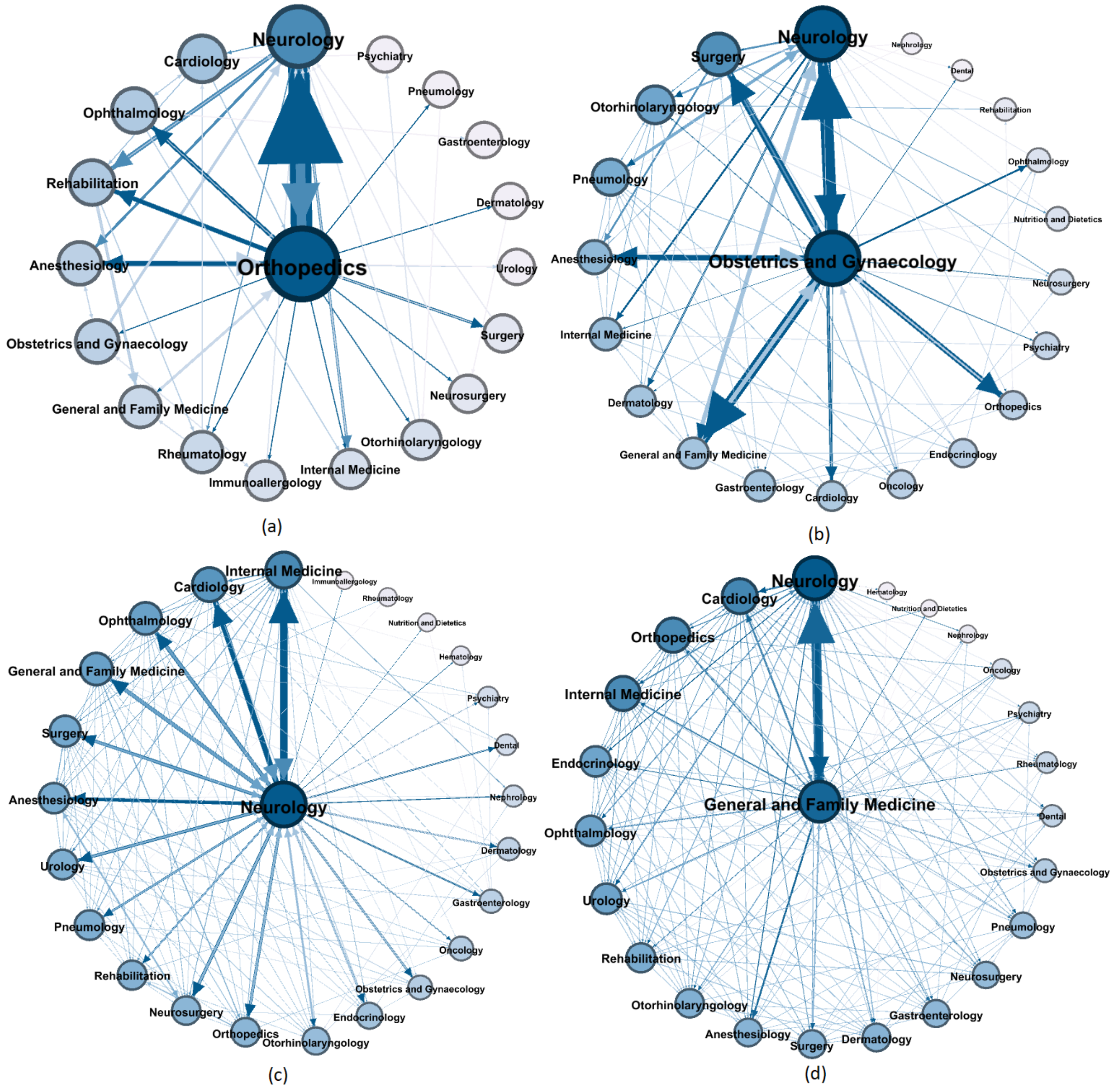
Cluster number	Jaccard median	Dice median	R. Rate median	Jaccard avg	Dice avg	R. Rate avg	Jaccard std	Dice std	R. Rate std
1	0.129	0.114	0.706	0.239	0.168	0.707	0.248	0.124	0.106
2	0.696	0.41	0.696	0.651	0.388	0.688	0.156	0.073	0.097
3	0.708	0.415	0.708	0.703	0.408	0.726	0.138	0.063	0.09
4	0.697	0.411	0.742	0.701	0.411	0.728	0.081	0.028	0.078
5	0.735	0.424	0.758	0.729	0.42	0.742	0.081	0.027	0.08
6	0.71	0.415	0.71	0.704	0.412	0.712	0.065	0.023	0.066
7	0.722	0.419	0.731	0.722	0.419	0.73	0.047	0.016	0.047
8	0.724	0.42	0.731	0.725	0.42	0.731	0.043	0.014	0.044
9	0.622	0.384	0.704	0.617	0.38	0.69	0.072	0.028	0.065
10	0.714	0.417	0.727	0.711	0.415	0.724	0.044	0.016	0.045
11	0.728	0.421	0.744	0.729	0.422	0.743	0.027	0.009	0.026
12	0.679	0.404	0.682	0.658	0.395	0.662	0.095	0.037	0.096

three correspondence measures, namely, Jaccard Index, Recovery Rate (R. Rate) and Dice Coefficient are calculated, whose results are presented in Table 4.5. The desire for these measures is that the median and average values be as close to unity as possible, while the standard deviation as close to 0. As already mentioned in [39], where the *Reuma.pt* database was used to study therapy switches, the median and average values for the Dice coefficient are smaller when compared to Jaccard and Recovery Rate, which are relatively high for every cluster. It is interesting to notice that, as the number of elements in the clusters increases, the average and median values of Jaccard and R. Rate tend to increase, while the corresponding standard deviations tend to decrease. Some of these values are not as satisfying as desired for considering the cluster stable. For instance, the first cluster does not have as satisfying results as the remaining clusters, being even considered less stable than cluster 9, which was identified as being composed by *outliers*. It was not expected that the *outliers* group would have such good stability measures since, for instance, the average SS of that cluster is close to 0 and a great part of its elements have a negative SS associated. Moreover, with the intuit of understanding which clusters could be considered most stable considering mainly the average and standard deviation measures, the mean value of each one of the indices' *avg* and *std* was obtained. Given these values, it was determined which clusters had a higher average Jaccard, Dice coefficient and Recovery Rate and a lower standard deviation than the correspondent mean across all clusters, which are highlighted in Table 4.5. Besides cluster 3, which is the one with the highest average SS, clusters 7, 8, 10 and 11 are the next four better formed groups, in terms of SS, which is in line with their stability.

In order to obtain a more visual look at the clusters formed, a directed graph was developed for each cluster, where nodes represent medical speciality consults existing within each cluster and edges serve as the number of patients who underwent a certain transition at least once. The size and color of nodes indicate the incidence of that medical consult within the cluster, while the thickness of the edges show whether a greater or smaller amount of patients in that cluster undergo that transition from one speciality to another. The nodes are displayed in a circular manner, ordered by the prevalence of the type of appointment.

As an example of the obtained graphs, Figure 4.19 represents the activity within clusters 3, 5, 11 and 12 where it is clear that Orthopedics, Obstetrics and Gynaecology, Neurology and GFM, respectively, are the dominant consults. Beyond the strong transition wise relation between the first consult from each cluster with Neurology consults, it is possible to see that:

- Patients who start their activity in Orthopedics, apart from Neurology, get sent to Ophthalmology, Rehabilitation, and Anesthesiology consults, more often than to other specialities present in cluster 3 and quite a reasonable amount of patients transition from Neurology to Rehabilitation and Anesthesiology. It is also possible to detect a stronger transition wise relation between Rehabilitation and GFM and subsequently to Orthopedics;



(a) (b) (c) (d)

Figure 4.19: Directed graphs showing the transitions between different medical consults within clusters (a) 3, (b) 5, (c) 11 and (d) 12, where Orthopedics, Ob-Gyn, Neurology and GFM consults prevail, respectively. Bigger and darker colored nodes indicate higher occurrence of that consult in the cluster, while wider edges indicate more patients underwent that transition at least once. The nodes are displayed in a circular manner, ordered by the prevalence of the type of appointment.

- The majority of patients who start off at Ob-Gyn consults, apart from Neurology, more recurrently carry on to Surgery, Anesthesiology, GFM and Orthopedics consults, comparatively to others. Not as often but still more recurrent than others, patients that form cluster 5 transition from Neurology to ORL, Pneumology, Internal Medicine and Dermatology and from GFM and Pneumology to Neurology;

- Regarding cluster 11, it is possible to identify transitions which stand out much more than others. For instance, it is common to find patients being redirected from Neurology to the top eleven most prevalent specialities after Neurology, being that the most prevailing ones are to Internal Medicine and Cardiology. Regarding incoming Neurology patients, these more often are being redirected from Internal Medicine, Cardiology, Ophthalmology and GFM. Transitions concerning other medical specialities besides the dominant one do not stand out in this subset;
- Finally, cluster 12, the one with the most patients, is seen as an heterogeneous one, transition wise. There are merely two transitions that stand out the most, which are transitions between GFM and Neurology consults. A slightly more significant amount of patients in this cluster can also be seen transitioning from GFM to Cardiology, Orthopedics, Internal Medicine and Anesthesiology, since these edges slightly stand out in the graph. Being the biggest sub-group of patients obtained, composed by 261 of them, it is natural to exist such diversity of transitions, remaining few of them that stand out. In addition, GFM is a heterogeneous medical speciality, justifying the fact that more specific patterns are not encountered.

4.2.2.D Clustering of female and male population with Dementia

So to evaluate how the clustering process would turn out in male and female patients with Dementia individually, the process of parameter optimization and clustering itself undergone for the whole population, was repeated for these two separate data sets.

From the 1116 and 749 female and male patients, respectively, 658 and 460 of them remained to be clustered resorting to the AliClu algorithm, having filtered all those who had attended more than twenty-one consults along their pathway, regardless of the 95 percentile registered for the number of consults correspondent to each data set, in order to guarantee that the same group of patients was being analysed.

Moving on to the parameter optimization process, the parameters analysed were the ones presented in Table 4.1, resorting to the same pipeline previously presented. Once again, a gap penalty of 0.1 and temporal penalty of 10 proved to be the best values for these parameters, for both subsets of patients. The optimum number of clusters was reached for k = 12 for the female subset and k = 10 for the male one.

By observation of Table 4.6, comparing as well with Table 4.3, it is possible to see that female patients alone were stratified in the same way as considering the whole population, while with male patients, a different stratification was obtained. Patients of both genders who initiate their pathway in GFM, Neurology, Internal Medicine, Anesthesiology, Cardiology and Surgery consults are subdivided in the same manner as for the whole set of Dementia patients. It was expected that patients with prevalence of Ob-Gyn consults in their pathway were only formed when dealing with female patients alone. However,

Table 4.6: Number of elements and prevailing medical speciality consult within each cluster for the female and male Dementia data sets.

Female dataset		Male dataset	
Cluster label (Number of elements)	Prevailing medical speciality consult	Cluster label (Number of elements)	Prevailing medical speciality consult
1 (10)	Endocrinology	1 (6)	Pneumology & Neurology
2 (13)	Nutrition and Dietetics	2 (11)	Pneumology
3 (16)	Orthopedics	3 (15)	Urology
4 (19)	Pneumology	4 (24)	Surgery
5 (33)	Ob-Gyn	5 (39)	Cardiology
6 (38)	Surgery	6 (48)	*Outliers*
7 (39)	Cardiology	7 (57)	Anesthesiology
8 (49)	Anesthesiology	8 (61)	Internal Medicine
9 (60)	*Outliers*	9 (87)	Neurology
10 (91)	Internal Medicine	10 (112)	GFM
11 (143)	Neurology	-	-
12 (147)	GFM	-	-

it is possible to see that male patients with prevalence of Endocrinology, Nutrition and Dietetics and Orthopedics were not grouped, which was not expected. On the other hand, the algorithm was able to gather a group of patients that were considered *outliers* before, which are the ones who have incidence of Urology consults in their history. Another interesting observation is that the first cluster formed for the male population gathers patients who have repetitive patterns of both Pneumology and Neurology consults instead of inserting these patients in clusters 2 and 9.

In order to understand if the choice of *k* would influence the subsets formed for the male population, the partitions obtained for male patients with *k = 11* were visually inspected. Note that *k = 12* would not make sense for this population, since a subset for male patients with prevalence in Ob-Gyn consults could not be obtained. However, the groups of patients missing from the ones formed with *k = 10* are still not formed. The eleventh cluster contains patients who initiated their hospital activity in GFM, resulting in two clusters formed by these patients, instead of one.

Summarizing, we can conclude that the male patients with Dementia are slightly less heterogeneous when it comes to their medical consult clinical pathway, when compared do females.

4.2.3 Phenotype screening per cluster

In order to attempt identification of certain patient characteristics and phenotype patterns within the different clusters obtained, a phenotype screening, similar to the one completed initially, was put in practice. Figure 4.20 represents the gender distribution for each of the twelve clusters. Overall, the female population dominates all clusters, which goes according to expectations, since the data set in

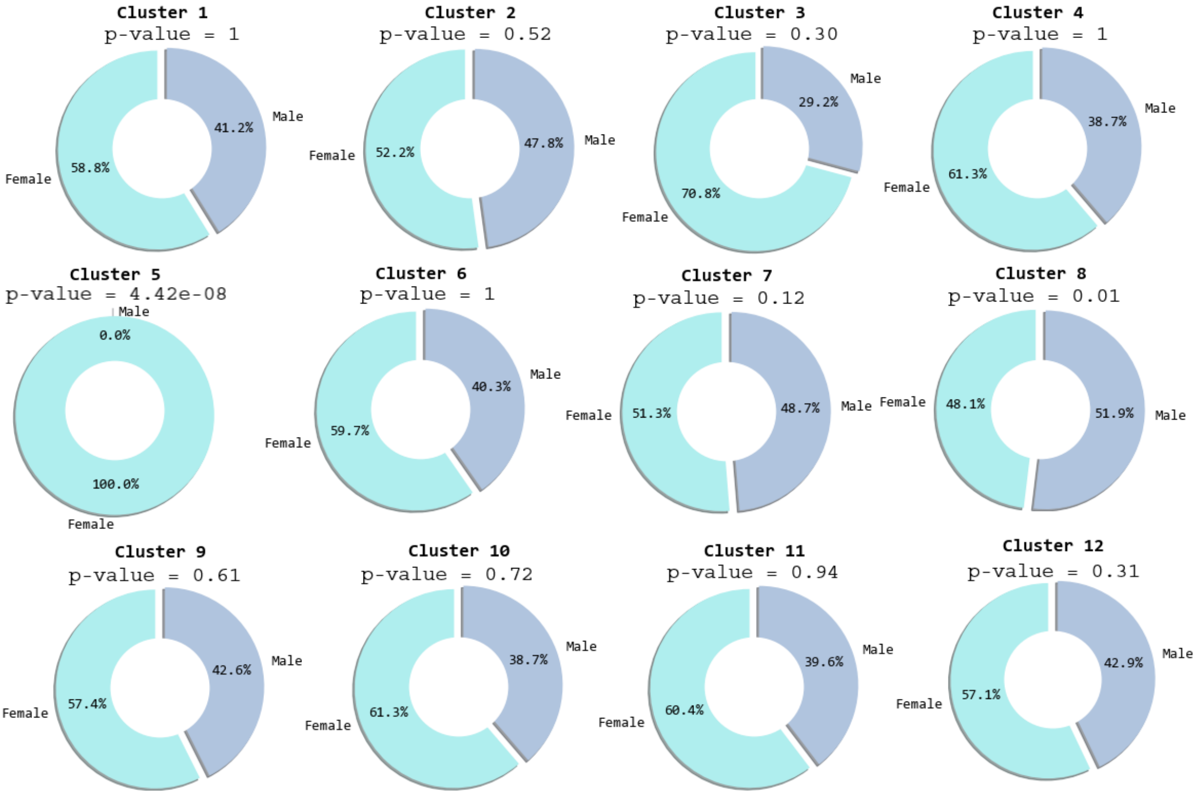


Figure 4.20: Gender distribution per cluster and relative p-values resultant from Fisher's exact test.

study contains a higher percentage of female patients. As expected, the patients in the fourth cluster are all female, since this is the cluster that groups patients with prevalence of Ob-Gyn consults in their history. Group 8, analogous to patients who initiate their hospital activity in Anesthesiology consults, presents a slight male dominance. In order to verify the significance of the gender proportions per cluster relatively to the overall data set, the p-value from Fisher's exact test was calculated. A p-value is considered significant below 0.05, meaning that the difference in female and male proportions in that cluster is significant and not due to the proportion of the whole set. Hence, only the Ob-Gyn and the Anesthesiology sub groups have a significant gender proportion, as expected, since the first one only contains females and the second has male dominance.

Furthermore, Figure 4.21 represents the mean and standard deviation of the number of consults attended by patients within each sub-group. The mean number of consults range only from approximately 5 to 9, however, the standard deviation values are relatively high for all clusters, indicating that the number of consults attended by patients in each cluster can have a high variance. Clusters 1, 3, 9 and 11 are the ones which contain patients who attended the least number of consults in the considered time window, while cluster 4 is formed by those who had more hospital visitations, having the highest average number, as well as the wider standard deviation.

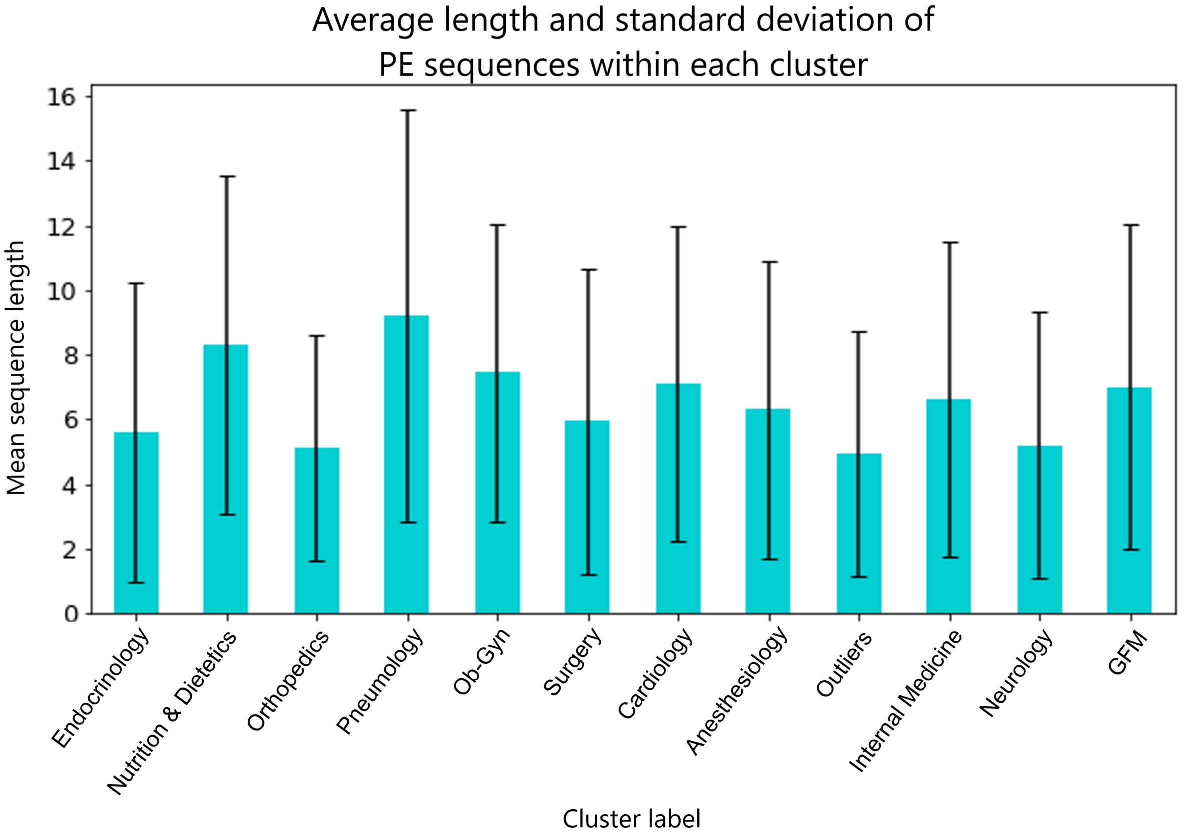


Figure 4.21: Mean and standard deviation of consult visitations by patients within each cluster formed for (g = 0.1, Tp = 10, k = 12)

Regarding the distribution of patients by age group, the results are presented in Figure 4.22. The initial age distribution obtained for the whole population indicates that the majority of Dementia patients are inserted in an age group between 60 and 100 years old, with some *outliers* between 20 and 60 years old. The majority of the sub-groups obtained are formed by patients which fit that same distribution, however, some stand out for having slight distinctions. For instance, cluster 5 contains younger patients when compared with the remaining. This set of patients represents the one completely formed by female patients that initiate their pathway in Ob-Gyn consults hence, it makes sense that it includes lower aged patients, resulting in a vaster age distribution. Furthermore, groups 3 and 4 are the ones that accommodate patients with a smaller variance in age. Respectively, there is a 23 and 27 age difference between the older and younger patients, while for cluster 5 this difference was 41. However, the cluster with more age discrepancy between patients is 12, with a 45 year old gap, which incorporates 261 patients who start-off at GFM.

Regarding the number of chronic illnesses that these patients suffer from, their distribution per cluster is presented in Figure 4.23. The initial distribution presented in Section 4.1, indicated that the Dementia population had between two and thirteen chronic diseases, whereas patients that did not fit this distri-

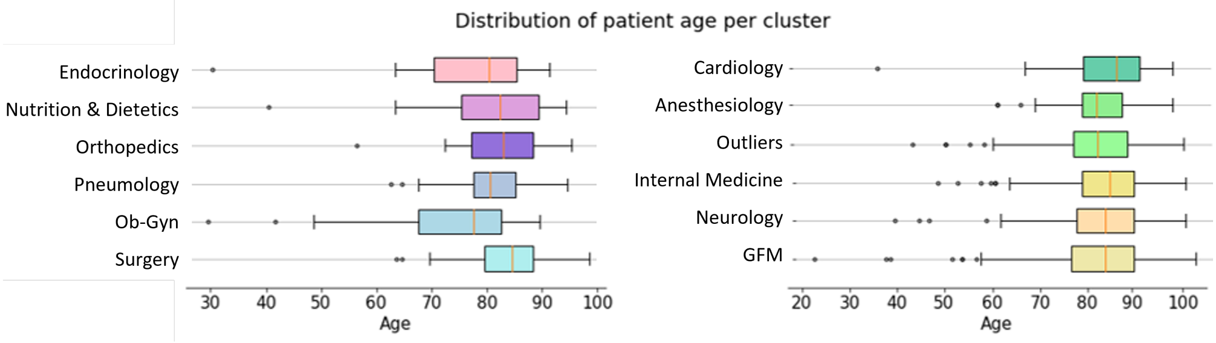


Figure 4.22: Age distribution per cluster.

bution could have up to twenty-three co-existing diseases. Additionally, the box plot obtained for the whole population showed that 50% of the patients had at most four chronic diseases. This finding only happens in cluster 3, since the remaining ones are composed by 50% of patients who suffer from at most five or six chronic diseases. The sub-group composed by patients who suffer from a wider range of chronic illnesses corresponds to the one that gathers patients who start off in a Pneumology consult (*i.e.*, cluster 4), which may have up to sixteen co-existing diseases. On the other hand, patients belonging to cluster 1, which gathers the ones who initiated their consult pathway in Endocrinology, suffer only between three and nine illnesses.

Furthermore, in order to understand the incidence of chronic diseases per partition obtained, the top four prevailing chronic diseases for each one of them were retrieved. Figure 4.24 represents the result of this analysis, where each set of bars represents the four most incident chronic diseases correspondent to each of the sub-groups obtained. With respect to the most prevailing chronic disease in each cluster, besides Dementia, it is possible to see that hypertension (HTN) is the one that dominates almost all sub-groups (ten out of twelve), with the exception of the first and third ones, where Type 2 Diabetes (T2DM) and osteoarthritis dominate, respectively. This is consistent with the fact that patients belonging to these clusters had regular presence in Endocrinology and Orthopedics consults, respectively. In addition, the

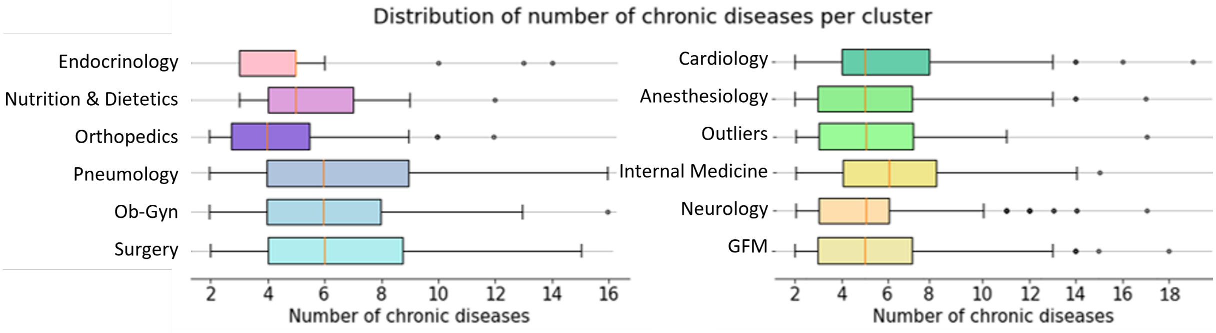


Figure 4.23: Number of chronic diseases distribution per cluster.

fourth most prevalent disease within the third sub-group is femoral neck pathologies, which is also an orthopedic disease.

Besides HTN, dyslipidemia is also a very common disease amongst these patients. As it is possible to see, it is present in the top four chronic diseases of ten of the total number of clusters, mostly as the second most incident, with exception of the orthopedics (C3), Ob-Gyn (C5) and cardiology (C7) ones, where it is the third or fourth most prevalent one. Moreover, a majority of seven out of twelve clusters are formed by a considerate amount of patients that suffer from cerebrovascular disease. However, it was already identified in the initial phenotype screening that HTN, dyslipidemia and cerebrovascular disease have high incidence in Dementia patients.

It is possible to identify specific chronic disease patterns that can be associated to the prevalence of certain medical specialities in each sub group. For instance, with respect to patients that belong to the pneumology cluster (C4), these have high prevalence of allergies and Chronic Obstructive Pulmonary Disease (COPD), which goes accordingly to their regular presence in pneumology consults, since allergies are very often associated to respiratory issues and COPD is a lung related disease.

Regarding the most frequent diseases experienced by patients that have their initial medical consult in Ob-Gyn (C5), which is 100% formed by females, we see that depression is the second most common

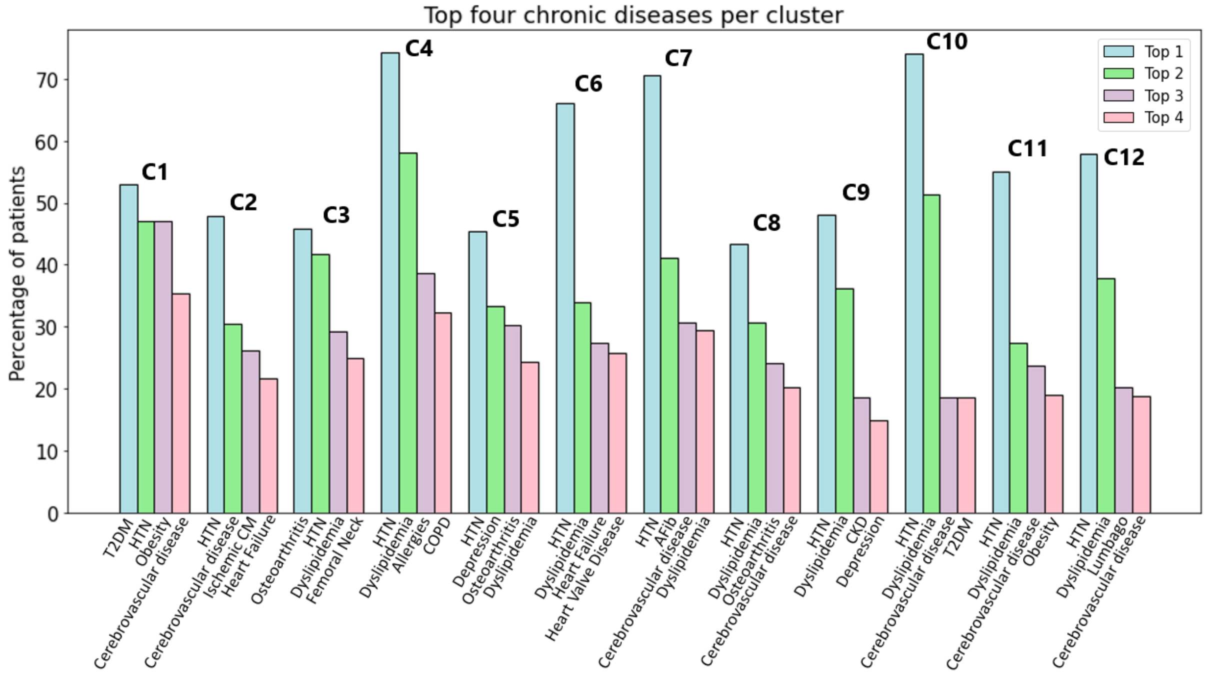


Figure 4.24: Incidence of the four most prevalent chronic diseases per cluster, excluding Dementia. Each set of four bars represents the incidence in each one of the sub-groups, which are sorted, from left to right, in ascending order. From blue, to green, purple and pink bars, the former represents the chronic illness suffered by most patients in the cluster, while the latter represents the one which is suffered by the least amount of individuals.

disease within them. Despite the fact that, as mentioned in Section 4.1, depression and Dementia have proven to be highly associated diseases, only this subset of patients have depression within the top four incident illnesses. Besides only being formed by female patients, it is also the sub-group that gathers younger patients, indicating that younger female patients with Dementia have much higher tendency of developing depression when compared to others. The fact that depression had more incidence in female Dementia patients was also previously identified upon the initial phenotype analysis presented in Section 4.1.

Now looking at the four most prevalent chronic diseases associated with a regular attendance to surgery consults (C6), besides HTN and dyslipidemia, these show higher tendency of suffering from

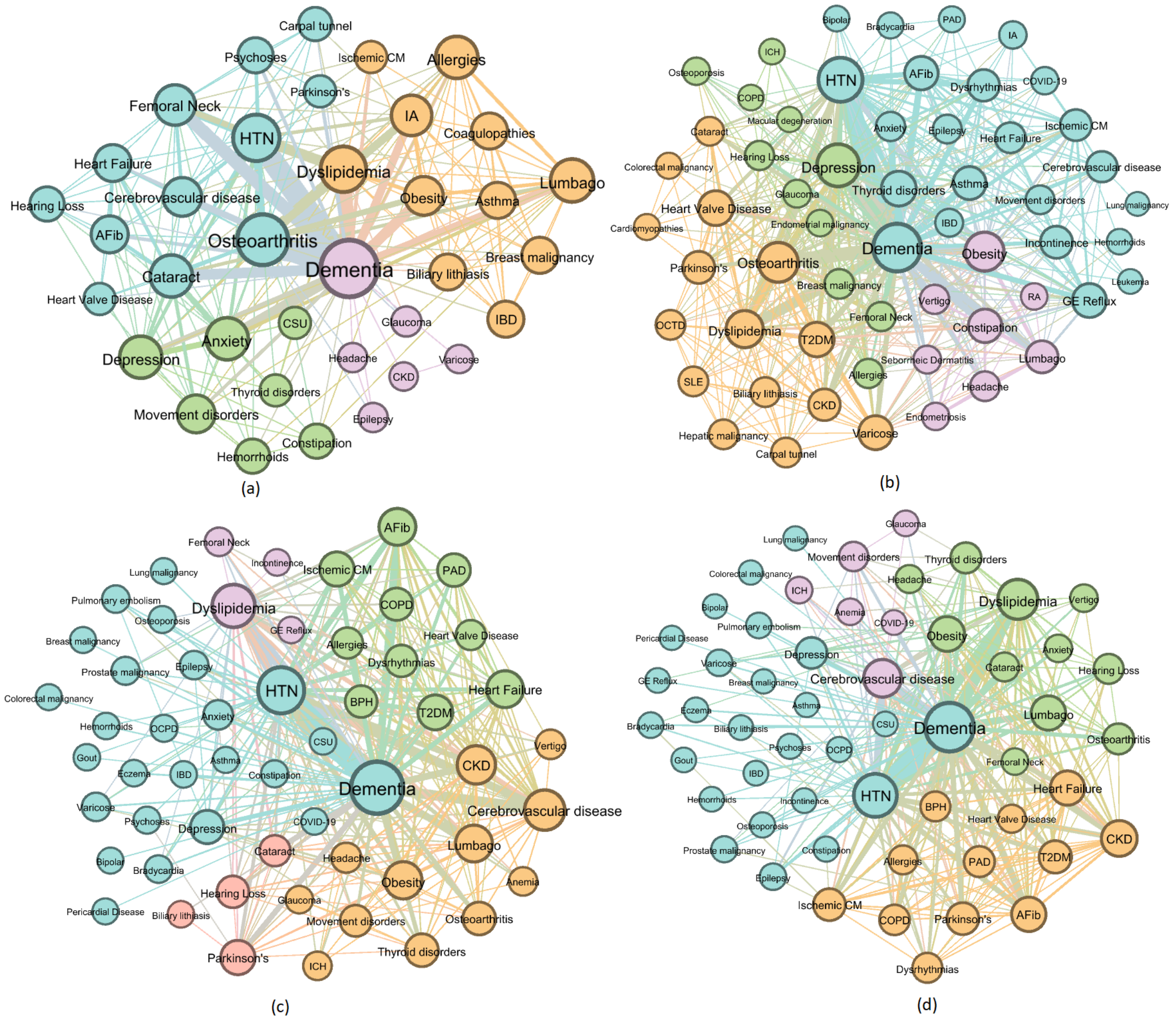


Figure 4.25: Graphs of chronic disease co-occurrence in patients belonging to clusters (a) three, (b) five, (c) eleven and (d) twelve, which presented regular presence in Orthopedics, Ob-Gyn, Neurology and GFM consults, respectively. Bigger nodes indicate higher prevalence of that disease in the subset, while wider edges indicate more frequency of co-occurrence of that pair of chronic illnesses.

heart failure and Heart Valve Disease, which may indicate that these diseases are more often related to surgical episodes when it comes to Dementia patients. Furthermore, the second most incident chronic illness amongst patients that form cluster number seven is AFib. This subset of the Dementia population was associated with a high attendance to Cardiology consults, which makes sense with the prevalence of this chronic disease. In the same line of thought, cerebrovascular disease, which is the third most existing chronic illness within this sub-group, is widely treated by Cardiologists [49], as well as Neurologists, which justifies the considerate number of patients suffering from this disease in the eleventh subset, which is the one with the higher prevalence of Neurology consults. Finally, lumbago, which was previously identified as also being very prevalent amongst these patients, is the third most common disease in patients that belong to cluster 12, which is formed by patients that start off in GFM consults.

In order to have a general overview of the incidence of the several chronic diseases suffered by patients belonging to each sub-group and not only the most common, co-occurrence graphs were obtained. Figure 4.25 presents these graphs corresponding to the orthopedics, Ob-Gyn, Neurology and GFM clusters. Once again, the modularity property was used to obtain the subgroups shown. In the first one, the prevalence of osteoarthritis, dyslipidemia and HTN, as seen in Figure 4.24, is clear. Femoral Neck pathologies, also included in the top four chronic diseases in this cluster, stands out practically as much as lumbago, allergies and depression, indicating that these illnesses are also quite prevailing. However, stronger co-occurrences happen between Dementia and the top four identified diseases. Besides these, patients with Dementia and IA, obesity, lumbago, cataract, anxiety, depression, are also quite common. The same goes for dyslipidemia and HTN, as well as for HTN and femoral neck pathologies. Regarding the Ob-Gyn sub set and the incident chronic diseases and their co-occurrences, it is possible to see the high prevalence of the top four identified chronic illnesses, which are HTN, depression, osteoarthritis and dyslipidemia. The more significant co-occurrences are also between these diseases, as well as between HTN and AFib and AFib and thyroid disorders. Similar patterns are detected regarding chronic diseases in the Neurology and GFM clusters where there is a clear dominance of the top four chronic conditions, presented in Figure 4.24, both in terms of prevalence and co-occurrence.

4.2.4 Medication analysis

Information on 35 263 prescriptions given to Dementia patients was made available for this analysis. Prescriptions given to patients who did not integrate any of the twelve obtained clusters were mapped out of the medication data set and the remaining medication data was mapped to each of the clusters. The medication data corresponding to the 1 118 that underwent the clustering process was isolated, being that all information presented from now on corresponds only to that fraction of patients.

A preliminary analysis was done to identify the percentage of patients who did not have any registered medication, as well as the average number of prescriptions and average number of different medication taken per patient, which results are presented in Table 4.7.

Table 4.7: Percentage of patients with no associated medication and average numbers of total and different prescriptions per patient.

Fraction of patients with zero registered medication	16.9%
Average number of prescriptions per patient	25
Average number of different medicines per patient	11

It is possible to see that quite a significant amount of patients do not take any medication. However, this has to be interpreted in a critical manner. The patient records gathered for this study are merely from HLL, which means that at least when attending a consult with a medical specialist in this facility, no medication was prescribed. However, it may not directly mean that these 16.9% of patients do not take any medication at all. The average number of prescriptions per patient considers all repetitions of prescriptions for the same medicine, while the number of different medicines considers only unique occurrences of a medicine in a patients' history. Being mostly a cohort composed by older patients, it is natural that these averages tend to be high, since, besides Dementia, as seen in the previous section, these patients suffer from a significant amount of additional comorbidities.

Furthermore, the average number of different medicines taken by patients in each cluster was also assessed. Patients belonging to clusters 3, 9 and 11, corresponding to the Orthopedics, *Outliers* and Neurology subgroups are the ones that showed the lowest medicine variety uptake, having demonstrated averages of seven, five and six, respectively. On the other hand, patients belonging to the Pneumology cluster (number 4), showed the highest diversity of medications, presenting an average of fourteen different ones per patient.

The next step was identifying the top 5 drugs prescribed to patients per cluster obtained. Results are presented in Figure 4.26. It is possible to establish some associations between the most prevalent medications in each cluster with the most incident medical consults and chronic diseases:

- The first cluster, Endocrinology, has prevalence of medication related to the treatment of diabetes, as well as the needle used by diabetic patients, which makes sense both with the most incident medical speciality, as well as with the high prevalence of type 2 diabetes in patients that form this subgroup.

- The nutrition and dietetics cluster has higher prevalence of vitamins and statins, which are usually prescribed to reduce cholesterol levels.

- The third cluster, correspondent to the Orthopedics one, has prevalence of calcium carbonate, which is used to enhance bone, muscle and nervous system health, as well as magnesium metamizole which is widely used to treat severe pain, which goes accordingly to the type of patients that form this subset, since orthopedics patients usually have to deal with a significant amount of pain.

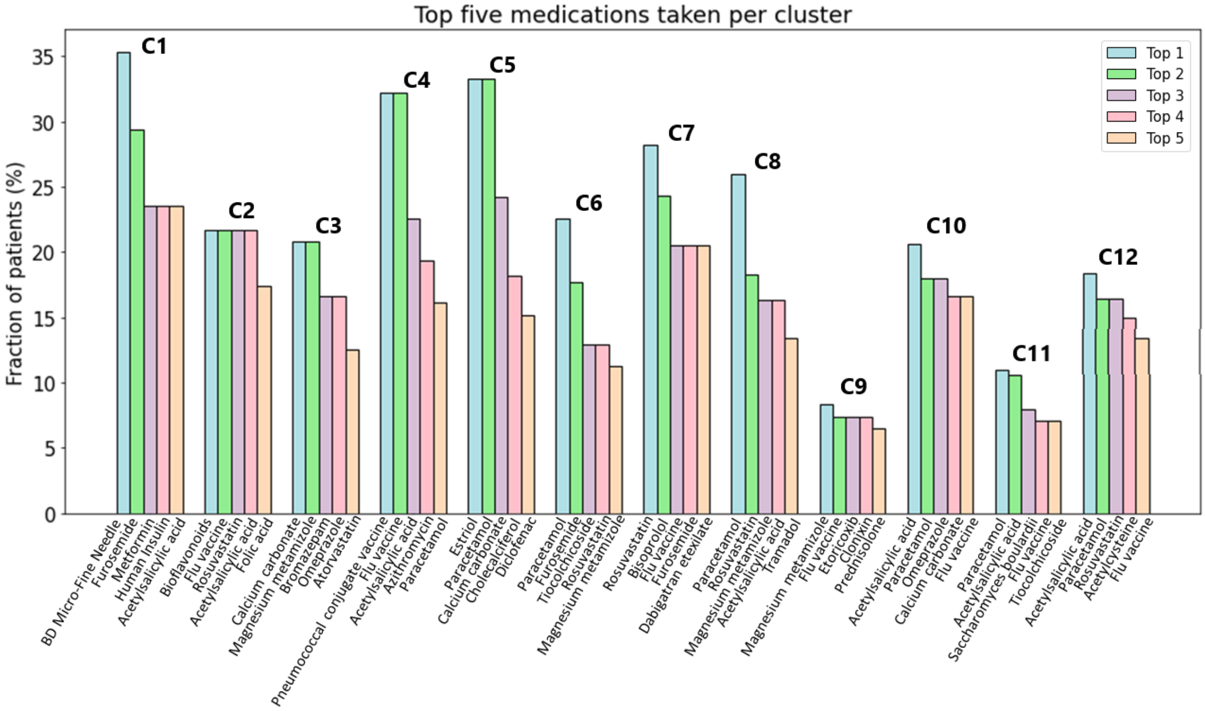


Figure 4.26: Top five medication taken by patients belonging to each one of the twelve clusters. Blue bars correspond to the most prevalent, followed by the green, purple, pink and orange bars as second, third, fourth and fifth most prevailing.

- The pneumococcal conjugate vaccine is the most prevalent medicine given to pneumology patients in C4, which is a vaccine to fight illnesses caused by pneumococcal bacteria, which normally cause diseases in the lungs.

- The Ob-Gyn cluster has estriol as most common medicine, which is a drug given to female patients, normally to treat symptoms related to menopause. Two types of vitamins are also recurrent, namely cholecalciferol and calcium carbonate.

- The cluster that gathers patients with prevalence of surgery consults has incidence of analgesic medication, such as paracetamol and thiocolchicoside, as well as rosuvastatin, whcih is a statin prescribed to reduce cardiac risk and furosemide which is a loop diuretic medication, used to prevent liquid build-up caused by heart failure, which represents the third most common chronic disease amongst this group of patients.

- The seventh cluster has rosuvastatin as the most common medicine and furosemide as the fourth most common, which goes accordingly to the fact that these are used to prevent cardiac problems, as mentioned in the previous item. Also included in the top five medications is bisoprolol, which is used to treat high blood pressure and dabigatran etexilate, which is prescribed to treat and avoid blood clots, which all go accordingly to the fact that this group gathers cardiac patients.

- The Anesthesiology cluster has included in its top five medicines analgesics and anti-inflammatory medication, such as paracetamol, tramadol and acetylsalicylic acid, also known as aspirin, as well as rosuvastatin, for cardiac issues.

- The ninth cluster, which is composed by *outliers*, has a more heterogeneous incidence of medications, due to the heterogeneity of the patients who form it.

For the remaining three clusters, it is not as immediate to establish relations between medication, prevalent consults and chronic diseases. These are the subgroups that gather patients with high recurrence in Internal Medicine, Neurology and GFM consults, formed by a relatively high amount of patients, which may justify the fact that the prevailing medication in these clusters are more generalist.

Moving on to the analysis of the prevalence of not medicines alone, but classes of medications, the results can be seen in Table 4.8. It includes only results with respect to medication classes that are common in more than five percent of the whole dataset.

It is possible to see that the most prevalent group within Dementia patients are analgesics. It has been previously noted that Dementia patients have quite a high tendency of suffering from chronic pain, which may be manageable but hard to communicate for these patients due to communication issues related to the disease. Achterberg et al. [50] states that more than half of the population with Dementia sense daily pain, which may be due to several reasons, including their comorbidities, such as cardiac or musculoskeletal diseases. It is also mentioned that patients that do experience pain are more likely to have a faster memory decline when compared to others, having also higher probability of developing other behavioral symptoms, such as depression, aggressive behavior, verbal abuse and wandering [50].

Statins are the second most common medication group, which are ones given to reduce cholesterol levels, which goes accordingly to the prevalence of dyslipidemia amongst Dementia patients, which is a disease, as previously mentioned, that elevates cholesterol levels. Furthermore, statins have also proven to have an influence in decreasing the chances of developing Dementia and Alzheimer's disease, having a positive impact in improving cognitive impairment [51].

Moving on to the third most prevalent medication class, Non-steroidal anti-inflammatory drugs (NSAIDs) are some times prescribed to reduce pain, which is in line with the prevalence of analgesics. Other uses include fever decrease, inhibition of blood clots and in some cases reduction of inflammation.

Diuretics are drugs that reduce blood pressure by making the kidneys release more sodium into the urine, which helps in removing water from blood vessels, decreasing the amount of fluid in circulation. A literature review on the impact of diuretics in cognitive impairment demonstrates an inverse relation between the incidence of Dementia and the use of these medications, more specifically, in patients that suffer from HTN and have a higher risk of developing Dementia [52]. Due to the high incidence of HTN

Table 4.8: Prevalence of twenty classes of medications in each cluster and in the whole data set.

Medications	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Dataset
Analgesics	31.25	28.57	45.0	37.93	43.33	43.14	30.88	46.07	16.46	38.93	25.95	34.18	33.8
Statins	25.0	47.62	15.0	37.93	23.33	33.33	42.65	30.34	10.13	36.64	24.05	38.82	31.65
NSAIDs	18.75	14.29	45.0	34.48	46.67	37.25	29.41	43.82	32.91	29.77	18.35	33.76	31.32
Diuretics	37.5	33.33	15.0	27.59	26.67	31.37	47.06	31.46	11.39	42.75	24.05	30.8	30.57
Steroids	37.5	19.05	20.0	37.93	36.67	37.25	39.71	25.84	31.65	34.35	20.89	30.8	30.25
Antidepressives	18.75	23.81	20.0	37.93	23.33	27.45	33.82	26.97	13.92	23.66	24.68	22.36	24.22
Benzodiazepines	18.75	38.1	25.0	27.59	16.67	23.53	32.35	23.6	18.99	22.9	18.35	23.63	23.04
Antiplatelets	37.5	28.57	5.0	31.03	10.0	15.69	26.47	20.22	8.86	29.01	22.15	23.21	21.96
Beta-blockers	25.0	23.81	10.0	20.69	20.0	21.57	36.76	24.72	12.66	19.85	18.35	16.46	19.91
ACE-ARBs	18.75	19.05	10.0	20.69	10.0	21.57	27.94	22.47	7.59	22.14	13.29	21.94	18.95
Vitamins	0.0	42.86	10.0	17.24	10.0	17.65	14.71	7.87	10.13	19.85	14.56	15.61	14.96
Anticoagulants	6.25	19.05	15.0	13.79	10.0	17.65	39.71	16.85	3.8	9.92	15.19	8.02	13.46
Calcium channel blockers	18.75	9.52	5.0	6.9	3.33	17.65	35.29	8.99	10.13	17.56	9.49	11.39	13.24
Bronchodilators	12.5	14.29	10.0	34.48	6.67	11.76	14.71	11.24	7.59	13.74	14.56	10.97	12.7
Anxiolytics	0.0	9.52	20.0	17.24	13.33	9.8	14.71	15.73	3.8	14.5	10.13	14.77	12.59
Antidiabetics	68.75	9.52	0.0	17.24	13.33	7.84	10.29	12.36	3.8	8.4	8.23	15.19	11.52
Dementia	12.5	19.05	0.0	17.24	3.33	15.69	13.24	6.74	5.06	14.5	15.19	9.7	11.3
Antipsychotics	6.25	4.76	10.0	6.9	0.0	5.88	8.82	7.87	3.8	9.16	9.49	8.86	7.86
Vasodilators	6.25	9.52	0.0	3.45	3.33	1.96	10.29	6.74	2.53	11.45	5.06	12.24	7.86
Thyroid medicines	0.0	4.76	0.0	6.9	10.0	9.8	8.82	10.11	3.8	6.11	5.06	7.59	6.78

in the considered dataset, it makes sense having a high prevalence of this class.

Steroids can be used for distinct purposes, which include treating diseases that lead to muscle loss, dealing with hormonal issues and blood vessel inflammation. Their high incidence is most probably related to the treatment and soothing of symptoms caused by Dementia and not the disease itself, since the use of steroids has been previously linked to memory and behavioral deficits [53].

Antidepressives and benzodiazepines are regularly used to deal with sleep disorders, chronic pain, anxiety and depression, which goes accordingly to the incidence of depression in these patients, as well as comorbidities that lead to pain and abnormal sleep such as osteoarthritis and lumbago. However, it was interesting to verify that some studies link this type of medication to an increased risk of Dementia, due to the fact that a long-term use may lead to reduced brain activity and accumulation of cognitive deficiencies, escalating the risk of Dementia in the elderly population [54] [55]. On the other hand, [56] reported that treatment with a type of antidepressives, which are the selective serotonin reuptake inhibitors (SSRIs), lead to cognitive improvements in Dementia patients. Given that almost all antidepressives considered for this group, shown in Table 3.4, are indeed SSRIs, except for the last four, which are not but have similar properties, their prevalence is most probably to control cognitive decay associated to Dementia. Anxiolytics and antipsychotics are also associated to the treatment of anxiety and depression. Antipsychotics are many times prescribed to control behavioral symptoms, such as agitation, aggression and psychosis in Dementia patients and proved effective [57].

Antiplatelet drugs are also quite common amongst these patients, which is in line with the fact that these are prescribed to reduce the risk of blood clots. Furthermore, these drugs have been linked to the prevention of VD. Chabriat et al. [58] states that the prevention of this type of Dementia passes by avoiding a stroke episode, which in turn involves managing its risk factors and use of antithrombotic medication, where antiplatelets are inserted. Moreover, it is also mentioned that after the diagnosis, the prescription of antiplatelet drugs is beneficial for the patients and may even decrease their risk of death [58]. Anticoagulants are also used to prevent blood clots and are many times prescribed to treat AFib, which is a common comorbidity amongst these patients, as previously signaled.

Beta-blockers, angiotensin converting enzyme inhibitors and angiotensin-receptor blockers (ACE-ARBs), calcium channel blockers and vasodilators are drugs prescribed for treatment of HTN. Due to the high incidence of this comorbidity in this Dementia cohort, it is natural that these types of drugs have a relatively high incidence. Nonetheless, Holm et al. [59] mentions previous studies that found a link between changes in blood pressure and risk of developing Dementia, explaining that low blood pressure has been pinpointed as a risk factor for developing Dementia, since it may cause declined cerebral activity, increasing the risk for VD. For this reason, anti-HTN medicines, which lower blood pressure, have been linked to an increased risk for cognitive and mental decline. On the other hand, studies involving ACE-ARB drugs and their influence in Dementia show that these have a positive impact in delaying or preventing the onset of the disease. Concerning calcium channel blockers, the literature is divided, having several studies associated these with a reduced risk of developing Dementia, while others associate them with an increased risk [60].

Medication used to treat dementia includes two types: (i) Cholinesterase inhibitors, which are prescribed to improve thought processes, such as memory, judgement, language and thinking and (ii) Glutamate regulators which, in turn, help improve memory, attention, language and the capability to perform a simple task. It is curious to see that the cluster with Neurology consult prevalence is not the one that includes more patients with these prescriptions. However, it is important to point out that neurologists treat a vast amount of diseases besides Dementia. It was expected that the prevalence of medication directed to treating Dementia symptoms would be higher, however it is only present in approximately eleven percent of the population. Nevertheless, it has been discussed above that other groups of medication are also aimed at treating or delaying cognitive impairment caused by this disease.

Finally, the incidence of vitamins, antidiabetics and thyroid medication is most probably due to the patients' age and the fact that T2DM and thyroid disorders are common amongst them.

Relatively to the prevalence of certain groups in the different clusters, it is possible to encounter some interesting patterns. The relatively high incidence in all clusters of beta-blockers and ACE-ARBs goes accordingly to the high tendency of these patients of suffering from HTN, heart failure and cerebrovascular disease. Beta-blockers, besides used for HTN treatment, also acts on abnormal heart rythms and it is interesting to verify that the cluster formed by cardiac patients, which is number seven, is precisely the one with higher incidence of this medication group. The same goes for ACE-ARBs, anticoagulants, calcium channel blockers. Furthermore, patients belonging to the Pneumology cluster (number four) are the ones that present higher uptake of bronchodilators, which makes sense, also with the fact that this specific subgroup showed high prevalence of chronic obstructive pulmonary disease.

4.2.5 Hospital admission and emergency analysis

Data on 2 078 hospital admissions (HA) and 9 991 emergency episodes (EE) was available for this part of the study. Once again, just as for the medication analysis, patients that did not belong to any of the obtained clusters were mapped out of the dataset, having pursued this analysis only with the hospital admission and emergency data belonging to the 1 118 patients that integrate the clustering process.

First of all, the fraction of these Dementia patients that actually had at least one of these occurrences and the average number of occurrences per patient was verified. In order to have a comparison baseline, the same data was obtained for the MM population in general. Results of this preliminary analysis are presented in Table 4.9.

Table 4.9: Percentage of MM and Dementia patients with hospital admissions and emergency episodes and corresponding average number of occurrences per patient, taking into consideration patients with zero occurrence, and corresponding ratios.

	Dementia patients	MM patients	Ratios Dementia/MM
Hospital admissions	46.24%	15.82%	2.92
Emergency episodes	83.27%	65.40%	1.27
Average number HA	1.86	0.24	7.75
Average number EE	8.94	3.06	2.92

Only 517 of the 1 118 considered Dementia patients (46.24%) were admitted to the hospital in the considered time frame, while a much higher number of 931 patients (83.27%) had an emergency episode. Regarding all patients with MM, 15.82% of them registered HA, indicating that Dementia patients had almost three times the tendency of being admitted when compared to the general population with MM. Considering EE, 65.40% of MM patients had emergency episodes, corresponding to 1.27 more emergency events when it comes to Dementia patients. Considering now the average number of occurrences per Dementia patient, approximately two HA were identified, while for emergency episodes that value was close to nine. For the MM population in general, an average of 0.24 HA and 3.06 EE were registered per patient. It is important to mention that these average values were calculated for both Dementia and MM populations, meaning that individuals with zero events were also considered for the calculations. We saw that the average number of HA for Dementia patients between January 2007 and August 2021 was 7.75 times higher than for MM patients, while the average EE was almost three times higher. This goes accordingly to what is encountered in the literature regarding emergencies and hospitalizations of Dementia patients. For instance, Shepherd et al. [61] reached the conclusion that people suffering from Dementia are more frequently hospitalised when compared to those without this disease. They found that the risk of being hospitalised is 1.42 times higher for Dementia patients when compared to non-Dementia patients and that hospitalization rates oscillate between 0.37 and 1.26 per patient, per year [61]. Reasons for this higher tendency for HA include the age factor, the high incidence of comorbidities and low independence and functional ability of these patients.

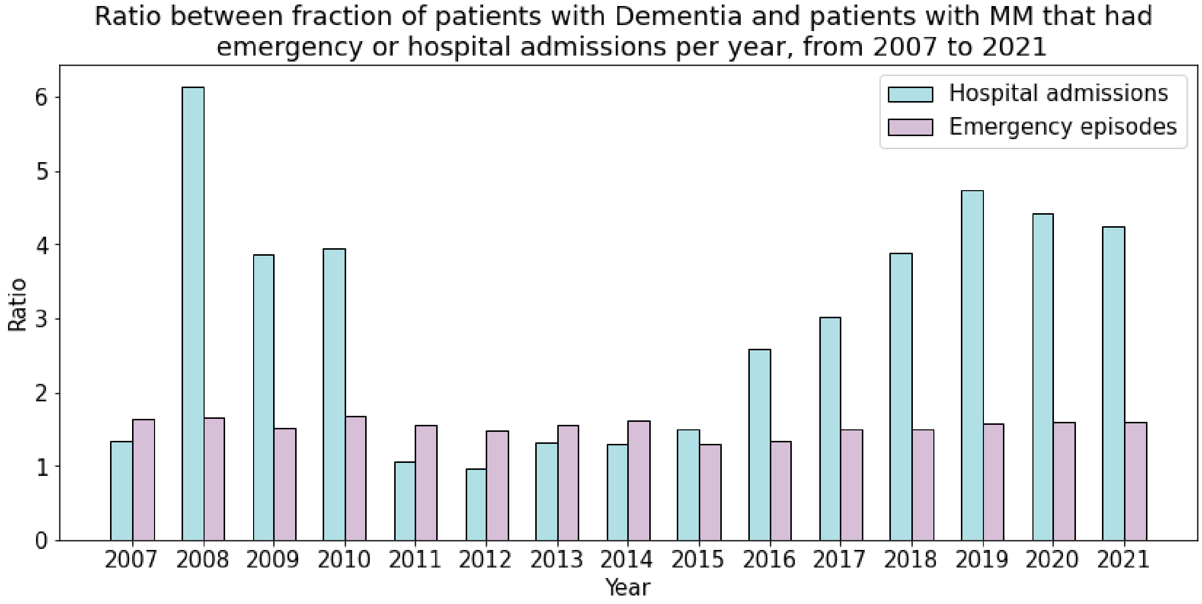


Figure 4.27: Fraction of Dementia patients that had hospital admissions or emergency events from 2007 to 2021, relative to the MM population.

Furthermore, in order to understand the tendency of these patients of having emergency events and requiring hospital admissions throughout the considered years, we did a survey of the portion of patients with registered emergencies and admissions per year and compared them with the same data for the MM population in general, having obtained their ratios, which are presented in Figure 4.27. A ratio higher than one indicates more tendency of Dementia patients having one of these occurrences, while one inferior to one illustrates the opposite.

It is clear that, whether considering hospital or emergency department admissions, Dementia patients have a higher tendency of having one of these occurrences when compared to the general MM population. While the data for EE is consistent throughout the years, for HA it is quite oscillating. Regarding emergencies, ratios vary between around 1.2 and 1.6, which means that patients with Dementia had 1.2 to 1.6 more probability of having an emergency episode throughout the years when compared to MM patients in general. This is consistent with the ratio obtained for our set of patients, presented in Table 4.9. With respect to hospital admissions, ratios can go from close to one up to six, depending on the year of activity. We believe these results should be critically interpreted, since the obtained values may be due to several factors. For instance, in 2008, according to hospital records, patients with Dementia were six times more probable of being admitted in contrast to the ones with MM. It was then verified that in this specific year, the percentage of Dementia patients with a HA in that year was particularly high and particularly low for MM patients. It is also important to mention the fact that there is much more data on emergencies than HA, which may influence the higher stability of these events throughout the years. Since the data on HA is not as much, a slight variance between Dementia and MM data will be much

Table 4.10: Number of patients in Cluster *i*, where *i = 1, 2, ..., 12*, that were admitted to the hospital (*a*), that were not admitted to the hospital (*b*), number of patients not in Cluster *i* that were (*c*) and were not (*d*) admitted to the hospital and corresponding odds ratio per cluster considered.

Cluster	a	b	c	d	Odds Ratio	p-value
Endocrinology	7	10	510	591	0.81	0.8
Nutrition & Dietetics	15	8	502	593	2.21	0.08
Orthopedics	8	16	509	585	0.57	0.2
Pneumology	13	18	504	583	0.84	0.7
Ob-Gyn	14	19	503	582	0.85	0.7
Surgery	39	23	478	578	2.05	0.008
Cardiology	43	35	474	566	1.47	0.1
Anesthesiology	56	48	461	553	1.40	0.1
Outliers	38	70	479	531	0.60	0.02
Internal Medicine	83	67	434	534	1.52	0.02
Neurology	105	122	412	479	1.00	1
GFM	96	165	421	436	0.60	0.0005

more significant.

Moving on to the assessment on how and if each cluster is associated to a higher or lower chance of being admitted to the hospital or to the emergency room, the odds ratio associated with each cluster and each of the events are presented in Tables 4.10 and 4.11, respectively. However, in order to verify if the odds ratio result is significant, it is important to assess the associated p-value. This means that an OR result is only meaningful if a p-value equal or lower than 0.05 is reached.

Focusing on the hospital admissions table, we can see that only the Surgery, the *Outliers*, Internal Medicine and GFM clusters have a meaningful odds ratio, since these are the only ones with a p-value lower than 0.05. Keeping in mind that an OR higher than one means that belonging to a certain cluster leads to higher outcome odds, we can conclude that patients who form the Surgery and Internal Medicine

Table 4.11: Number of patients in Cluster *i*, where *i = 1, 2, ..., 12*, that had an emergency episode (*a*), that did not have an emergency (*b*), number of patients not in Cluster *i* that did (*c*) and did not (*d*) have an emergency episode and corresponding odds ratio per cluster considered.

Cluster	a	b	c	d	Odds Ratio	p-value
Endocrinology	14	3	917	184	0.94	1
Nutrition & Dietetics	21	2	910	185	2.13	0.4
Orthopedics	20	4	911	183	1.00	1
Pneumology	28	3	903	184	1.90	0.5
Ob-Gyn	25	8	906	179	0.62	0.2
Surgery	55	7	876	180	1.61	0.3
Cardiology	67	11	864	176	1.24	0.6
Anesthesiology 8	85	19	846	168	0.89	0.7
Outliers	72	36	859	151	0.35	8.4e-06
Internal Medicine	131	19	800	168	1.45	0.2
Neurology	179	48	752	139	0.69	0.05
GFM	234	27	697	160	1.99	0.001

clusters have higher tendency of being admitted to the hospital. On the other hand, patients considered as *Outliers* and those who form the General and Family Medicine subgroup have lower associated odds of being admitted. The remaining odds ratio results are not significant since a meaningful p-value was not obtained.

Regarding Table 4.11 and following the same line of thought, only the OR results corresponding to the *Outliers*, Neurology and GFM clusters have a significant meaning. According to the OR values, patients belonging to the GFM subgroup have higher tendency of having an emergency episode, while the remaining two have lower associated odds of that same outcome.

5

Conclusions and future work

Contents

The work developed in this Thesis integrates a project developed by Hospital da Luz Lisboa, named IntelligentCare. This research aims at promoting a patient centered model to help manage patients with multimorbidity through analysis of electronic medical records using analytical methods. Specifically, my contribution with the project presented in this document, was to create a pipeline to identify patterns within Dementia patients, focusing on stratifying them based on their hospital activity. To this end, longitudinal data from HLL was used and two main approaches were explored, namely Markov chains and a temporal sequence alignment clustering process. Subsequently to identifying interaction patterns, each established subgroup was overseen in terms of demographic, complexity and risk factors.

The growing impact of multimorbidity in the population, coupled to the complexity surrounding Dementia patients and the fact that the incidence of this disease will tend to increase as the population lives longer, was the main motivation for this work. Focusing on a single disease, not considering the whole spectrum of a patient was identified as a problem in the care providing system. Scientific literature revealed that there is still a long way to achieve a patient-centered integrated care and we believe that analysing complex heterogeneous groups of patients, as is the case of Dementia ones, is an important first step in this direction. All the tools used and developed in this work can be easily adapted to explore other patient cohorts, which may help guide health providers in the best direction to provide treatment.

This chapter outlines the main remarks, the limitations encountered along the way and, finally, possible future work to be addressed.

5.1 Conclusions and limitations

Data quality and consistency in this type of studies is key for their success. The fact that electronic medical records may be biased or have insertion errors and that sometimes not all of a patients' information is available in a single EMR of a single health facility represent barriers. At the same time, detecting these errors or missing values can be a very time consuming and ineffective task. Developing a tool that could easily detect missing values or correct errors upon data insertion could bring many benefits to the effort of extracting meaningful information and patterns amongst EMRs.

The initial phenotype screening implemented for the whole cohort allowed to identify important characteristics regarding these patients, specially in terms of age group and additional chronic diseases. The literature indicated that the majority of Dementia patients integrate an age group higher than sixty-five years old, which was verified. However, onset ages that are not so common were also identified in this cohort, which was interesting to analyse. Several chronic diseases were associated to an increased risk of Dementia and some patterns were detected when comparing the incidence of certain chronics conditions in the Dementia population with their incidence in the MM population in general. In addition, it was also possible to recognize distinct prevalence when focusing on patients per gender. However, it has to be taken into consideration that some data may have been wrongly inserted. For instance, it is possible but very rare to have patients being diagnosed with Dementia below the age of forty and it is difficult to recognize mistakes from the EMR. The onset Dementia age analysis may also be dubious due to the fact that a patient may be diagnosed at a certain age but started developing symptoms prior to that diagnosis. Hence, this phenotype analysis pipeline developed in this work would probably benefit from a review and information extraction of clinical notes.

In line with the objectives that were set for this dissertation – the identification of patterns within Dementia patients regarding clinical pathways – two distinct tools were used. The clinical pathway analysis regarding medical consults attended by these patients allowed to identify which medical specialties this subset of MM patients face the most and relate them to the most prevailing chronic diseases. The Markov Chains approach allowed to pinpoint the most common transitions between medical specialities, considering the source consult, through development of transition matrices. Considering two distinct scenarios to obtain these matrices, one of them lead to the conclusion that follow up consults in the same speciality normally occur prior to moving to a different one. The second scenario allowed to identify patterns only within different specialities. Transition matrices were also obtained per gender, which allowed to understand the activity progression of these patients per gender, verifying which transitions were most frequent in each subgroup.

In order to attempt finding distinct subgroups with similar activity characteristics, AliClu clustering algorithm was applied to the dataset. Some limitations regarding the algorithm and its parameters optimization were encountered. Hence, adaptations were done in order to smooth the parameter optimization process to make it more efficient. However, due to the high complexity of the algorithm itself, it was only possible to test parameters until a certain point, so that it was still possible to run it in an acceptable running time. AliClu comprises a sequence alignment and a clustering phase. The first one compares each pair of sequence existent in the data set, both in terms of events as well as elapsed time between consecutive events, which is very time consuming. This may be a limiting aspect of using this pipeline in bigger data sets, since an increase in the number of sequences to work with will increase the complexity of the algorithm and its running time, making it ineffective to use. A careful revision of the implementation would be beneficial to make AliClu even more promising in this area.

Posterior to the parameter optimization process, twelve distinct clusters were obtained for Dementia patients regarding their medical consult activity. Similar patterns were found within each subset but the one that stood out more in each one was the first (and most prevailing) registered appointment, which all patients per cluster had in common. By resorting to directed graphs, it was possible to visualize the most prevalent clinical pathways regarding Dementia patients, per cluster formed, from which conclusions were drawn concerning most visited medical speciality consults, as well as most common transitions. Quantitative validation of these groups indicated that some were more stable than others, but in all of them it was possible to identify specific patterns.

Besides identifying medical consult patterns among each cluster, we also considered important to determine phenotype heterogeneity between sub groups. To that end, a phenotype screening was accomplished per subset, concerning gender, number of consults attended, chronic diseases, medication and, finally, hospital admissions and emergency episodes. It was interesting to observe that certain chronic diseases are prevalent, no matter the subset in which a patient is inserted. However, some clusters had incidence of specific chronic conditions that allowed to build associations with their most recurrent medical speciality consults. This has potential to, in the future, be used as a support decision tool for a health provider in deciding on the best course of treatment, taking into consideration all of a patients illnesses.

Results of the medication analysis allowed to pinpoint incidences of certain medication classes in certain clusters, being also possible to relate them to the most prevailing chronic diseases in that cluster. However, due to the amount and heterogeneity of medications taken by these patients, this analysis was complex and may even be biased. For instance, we were given data on prescriptions since the patients' beginning of treatment in the hospital in question. However, other prescriptions may have been given prior to that which are not taken into consideration.

Regarding the outcomes of the hospitalization and emergencies analysis, an evolution of the recurrence of these events throughout the years from 2007 to 2021 was done. With respect to the emergency events, these proved to be consistent, however, hospital admissions demonstrated quite a lot of variance. Nonetheless, we have to keep in mind that the amount of information data available on emergencies was much higher than the one on inpatient stays, which also has a big impact on the observed results.

Overall, it is possible to say that the pipeline applied to the Dementia cohort in question allowed to pinpoint characteristics and patterns regarding these patients. In the future, it may represent an important support tool for treating a patient as a whole and not only centered on a specific disease.

5.2 Future work

Relatively to the phenotype screening process, it could benefit from additional information. For instance, having access to written clinical notes that would contain extra information to be analysed, allowing to specify outcomes even further. Resorting to natural language processing tools would allow to gather more information on patients symptoms, supporting the recognition of the stage and severity of a certain illness. Data on chronic diseases besides their identification and diagnostic dates could also help to understand their onset phase and would represent an extra classification of patients. For instance, it would help to identify symptoms and occurrences that posteriorly lead to the diagnosis, allowing early signaling in the future. Other possible approaches could include exploring laboratory results that would allow learning common levels of certain blood components that would have significance in the disease in question.

In terms of possible improvements concerning the clinical pathway analysis, we believe that the clustering algorithm has potential to stratify patients taking into account temporal sequences, however, it would benefit from some adjustments in order to be viable to use in larger data sets. Furthermore, it would be beneficial to test other algorithms and techniques to stratify patients based in temporal sequences, in order to validate the stratification obtained with AliClu. Examples of some of these algorithms are Biclustering, PrefixSpan, CM-Spade and Fournier08.

Regarding medication, hospitalization and emergencies, we believe this analysis could benefit from additional information, such as indication of the specific reason for prescribing that specific medication and the causes of hospital admission and emergency event. An additional approach to studying the medication prescribed to these patients would be doing a survey of the ones prescribed for a same reason and investigating patient outcomes, understanding which stand as more or less efficient.

To finalize, in order to validate the whole pipeline of pattern identification developed in this work, it is important to apply it to different data sets, for instance considering a different disease or a specific age group. This would also lead to acquirement of important knowledge on a vaster amount of patients.

Bibliography

[1] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, “ Big data in healthcare: management, analysis and future prospects,” *Journal of Big Data*, vol. 6, no. 54, 2019.

[2] M. Rijken, V. Struckmann, I. Van der Heide, A. Hujala, F. Barbabella, E. Van Ginneken, and F. Schellevis, “How to improve care for people with multimorbidity in europe?” European Observatory on Health Systems and Policies - Policy Brief no. 23, 2017.

[3] Health IT Buzz, “EMR vs EHR – What is the Difference?” 2011. [Online]. Available: https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/emr-vs-ehr-difference

[4] J. M. Banda, M. Seneviratne, T. Hernandez-Boussard, and N. H. a. Shah, “Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models.” *Annual review of biomedical data science*, vol. 1, pp. 53–68, 2018.

[5] A. Shinozaki, “Electronic Medical Records and Machine Learning in Approaches to Drug Development.” *Artificial Intelligence In Oncology Drug Discovery And Development.*, 2020.

[6] V. Agarwal, P. Lependu, T. Podchiyska, R. Barber, M. Boland, G. Hripcsak, and N. Shah, “Using narratives as a source to automatically learn phenotype models,” in *Workshop on Data Mining for Medical Informatics*, 2014.

[7] Y. Halpern, Y. Choi, S. Horng, and D. Sontag, “Using anchors to estimate clinical state without labeled data,” in *AMIA Annual Symposium Proceedings*, vol. 2014. American Medical Informatics Association, 2014, p. 606.

[8] J. C. Ho, J. Ghosh, S. R. Steinhubl, W. F. Stewart, J. C. Denny, B. A. Malin, and J. Sun, “Limestone: High-throughput candidate phenotype generation via tensor factorization,” *Journal of biomedical informatics*, vol. 52, pp. 199–211, 2014.

[9] J. Almirall and M. Fortin, “The Coexistence of Terms to Describe the Presence of Multiple Concurrent Diseases.” *Journal Of Comorbidity, 3(1)*, pp. 4–9, 2013.

[10] C. Violan, Q. Foguet-Boreu, G. Flores-Mateo, C. Salisbury, J. Blom, M. Freitag, L. Glynn, C. Muth, and J. M. Valderas, “Prevalence, Determinants and Patterns of Multimorbidity in Primary Care: A Systematic Review of Observational Studies.” *Plos ONE, 9(7)*, 2014.

[11] R. Navickas, V. Petric, A. Feigl, and M. Seychell, “Multimorbidity: What Do We Know? What Should We Do?” *Journal Of Comorbidity, 6(1)*, pp. 4–11, 2016.

[12] A. Hassaine, D. Canoy, and J. S. et al., “Learning multimorbidity patterns from electronic health records using Non-negative Matrix Factorisation,” *Journal of Biomedical Informatics*, 2020.

[13] K. Wikström, J. Lindström, K. Harald, M. Peltonen, and T. Laatikainen, “Clinical and lifestyle-related risk factors for incident multimorbidity: 10-year follow-up of Finnish population-based cohorts 1982–2012.” *European Journal Of Internal Medicine, 26(3)*, pp. 211–216, 2015.

[14] Rijken, M., Struckmann, V., Van der Heide, I., Hujala, A., Barbabella, F., Van Ginneken, E. and Schellevis, F., “How to improve care for people with multimorbidity in Europe?” European Observatory on Health Systems and Policies - Policy Brief no. 23, 2017. [Online]. Available: https://www.euro.who.int/__data/assets/pdf_file/0004/337585/PB_23.pdf

[15] The Academy of Medical Sciences., “Multimorbidity: a priority for global health research.” 2021. [Online]. Available: https://acmedsci.ac.uk/policy/policy-projects/multimorbidity

[16] S. Ng, R. Tawiah, M. Sawyer, and S. P., “Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis.” *International Journal of Epidemiology, 47(5)*, p. 1687–1704, 2018.

[17] A. Hassaine, G. Salimi-Khorshidi, D. Canoy, and K. Rahimi, “Untangling the complexity of multimorbidity with machine learning,” *Mechanisms of Ageing and Development*, vol. 190, p. 111325, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047637420301214

[18] J. Zhou, F. Wang, and J. H. an Jieping Ye, “From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records,” *KDD ’14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 135–144, 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2623330.2623711

[19] W. H. Organization, “Dementia,” https://www.who.int/news-room/fact-sheets/detail/dementia, accessed: 2021-09-28.

[20] OECD, *Health at a Glance 2019*, 2019. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/4dd50c09-en

[21] F. Bunn, A.-M. Burn, C. Goodman, G. Rait, S. Norton, L. Robinson, J. Schoeman, and C. Brayne, "Comorbidity and dementia: a scoping review of the literature," *BMC Medicine*, vol. 12, 10 2014.

[22] B. Poblador-Plou, A. F. Calderón-Larrañaga, J. Marta-Moreno, J. Hancco-Saavedra, A. Sicras-Mainar, M. Soljak, and A. Prados-Torres, "Comorbidity of dementia: a cross-sectional study of primary care older patients," *BMC Psychiatry*, vol. 14, 03 2014.

[23] J. Ryan, P. Fransquet, J. Wrigglesworth, and P. Lacaze, "Phenotypic heterogeneity in dementia: A challenge for epidemiology and biomarker studies," *Front Public Health*, vol. 6, 2018.

[24] S. Verdi, A. F. Marquand, J. M. Schott, and J. H. Cole, "Beyond the average patient: how neuroimaging models can address heterogeneity in dementia," *Brain*, 04 2021, awab165.

[25] A. Y. Lee, "Vascular Dementia," *Chonnam Medical Journal*, vol. 47, no. 2, pp. 66–71, 2011.

[26] T. F. Outeiro, D. J. Koss, D. Erskine, L. Walker, M. Kurzawa-Akanbi, D. Burn, P. Donaghy, C. Morris, J.-P. Taylor, A. Thomas, J. Attems, and I. McKeith, "Dementia with Lewy bodies: an update and outlook." *Molecular neurodegeneration*, vol. 14, no. 1, 2019.

[27] A. L. Byers and K. Yaffe, "Depression and risk of developing dementia." *Nature reviews. Neurology*, vol. 7, no. 6, pp. 323–331, 2011.

[28] V. Camus, H. Kraehenbühl, M. Preisig, C. J. Büla, and G. Waeber, "Geriatric depression and vascular diseases: what are the links?" *Journal of affective disorders*, vol. 81, no. 1, pp. 1–16, 2004.

[29] Z. Huang, W. Dong, L. Ji, C. Gan, X. Lu, and H. Duan, "Discovery of clinical pathway patterns from event logs using probabilistic topic models," *Journal of Biomedical Informatics*, vol. 47, pp. 39–57, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046413001445

[30] R. Lenz and M. Reichert, "It support for healthcare processes – premises, challenges, perspectives," *Data Knowledge Engineering*, vol. 61, no. 1, pp. 39–58, 2007, business Process Management. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X06000784

[31] S. Wakamiya and K. Yamauchi, "What are the standard functions of electronic clinical pathways?" *International Journal of Medical Informatics*, vol. 78, no. 8, pp. 543–550, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505609000409

[32] F. ren Lin, L. shih Hsieh, and S. mei Pan, "Learning clinical pathway patterns by hidden markov model," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 2005, pp. 142a–142a.

[33] W. Michalowski, S. Wilk, A. Thijssen, and M. Li, “Using a bayesian belief network model to categorize length of stay for radical prostatectomy patients.” *Health care management science*, vol. 9, no. 4, pp. 341–348, 2006.

[34] M. G. Omran, A. P. Engelbrecht, and A. Salman, “An overview of clustering methods,” *Intelligent Data Analysis*, vol. 11, no. 6, pp. 583–605, 2007.

[35] M. Liao, Y. Li, F. Kianifard, E. Obi, and S. Arcona, “Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis.” *BMC Nephrology*, vol. 17, no. 25, 2016.

[36] R. Xu and D. C. n. Wunsch, “Clustering algorithms in biomedical research: a review.” *IEEE reviews in biomedical engineering*, vol. 3, pp. 120–154, 2010.

[37] A. Giannoula, A. Gutierrez-Sacristán, Álex Bravo, F. Sanz, and L. I. Furlong, “Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study.” *Scientific Reports*, vol. 8, 2018.

[38] M. Huang, N. D. Shah, and L. Yao, “Evaluating global and local sequence alignment methods for comparing patient medical records.” *BMC medical informatics and decision making*, vol. 19, 2019.

[39] K. Rama, H. Canhão, A. M. Carvalho, and S. Vinga, “Aliclu - temporal sequence alignment for clustering longitudinal clinical data,” *BMC Medical Informatics and Decision Making*, vol. 19, 2019. [Online]. Available: https://doi.org/10.1186/s12911-019-1013-7

[40] H. Syed and A. K. Das, “Temporal needleman-wunsch,” *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–9, 2015.

[41] Statistics.Com: Data Science, Analytics Statistics Courses, “Ward´S Linkage - Statistics.Com: Data Science, Analytics Statistics Courses,” 2021. [Online]. Available: https://www.statistics.com/glossary/wards-linkage/

[42] S. Zhao, J. Sun, K. Shimizu, and K. Kadota, “Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results,” *Biological procedures online*, vol. 20, 2018.

[43] T. Roeschl, “Fisher’s exact test from scratch with python,” Feb 2020. [Online]. Available: https://towardsdatascience.com/fishers-exact-test-from-scratch-with-python-2b907f29e593

[44] M. Szumilas, “Explaining odds ratios.” *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 19, pp. 227–229, 2010.

[45] "Alzheimer's disease genetics fact sheet," 2019. [Online]. Available: https://www.nia.nih.gov/health/alzheimers-disease-genetics-fact-sheet

[46] A. Y. Wong, J. Karppinen, and D. Samartzis, "Low Back Pain In Older Adults: Risk Factors, Management Options And Future Directions," *Scoliosis And Spinal Disorders*, vol. 12, no. 1, 2017.

[47] M. Papaleontiou and M. R. Haymart, "Approach To And Treatment Of Thyroid Disorders In The Elderly," *Medical Clinics Of North America*, vol. 96, no. 2, pp. 297–310, 2012.

[48] S. E. Jorgensen and B. Fath, *Encyclopedia of Ecology*. Elsevier Science, Jul 2008. [Online]. Available: https://www.123library.org

[49] "Neurology," Jun 2021. [Online]. Available: https://www.mayoclinic.org/departments-centers/cerebrovascular-diseases-critical-care-group/overview/ovc-20443614

[50] W. Achterberg, S. D. Lautenbacher, B. Husebo, A. Erdal, and K. Herr, "Pain in dementia." *Pain reports*, vol. 5, 2020.

[51] B. G. Schultz, D. K. Patten, and D. J. Berlau, "The role of statins in both cognitive impairment and protection against dementia: a tale of two mechanisms." *Translational neurodegeneration*, vol. 7, 2018.

[52] T. DeLoach and J. Beall, "Diuretics: A possible keystone in upholding cognitive health." *The mental health clinician*, vol. 8, pp. 33–40, 2018.

[53] "Reversible dementia from corticosteroid therapy." [Online]. Available: https://www.consultant360.com/articles/reversible-dementia-corticosteroid-therapy

[54] Y.-C. Wang, P.-A. Tai, T. N. Poly, M. M. Islam, H.-C. Yang, C.-C. Wu, and Y.-C. J. Li, "Increased risk of dementia in patients with antidepressants: A meta-analysis of observational studies." *Behavioural neurology*, 2018.

[55] Q. He, X. Chen, T. Wu, L. Li, and X. Fei, "Risk of dementia in long-term benzodiazepine users: Evidence from a meta-analysis of observational studies." *Journal of clinical neurology*, vol. 15, pp. 9–19, 2019.

[56] J. Schmitt, M. Wingen, J. Ramaekers, E. Evers, and W. Riedel, "Serotonin and human cognitive performance," *Current pharmaceutical design*, vol. 12, no. 20, pp. 2473–2486, 2006.

[57] R. R. Tampi, D. J. Tampi, S. Balachandran, and S. Srinivasan, "Antipsychotic use in dementia: a systematic review of benefits and risks from meta-analyses." *Therapeutic advances in chronic disease*, vol. 7, pp. 229–245, 2016.

[58] H. Chabriat and M. Bousser, “Vascular dementia: Potential of antiplatelet agents in prevention.” *European Neurology*, vol. 55, no. 2, pp. 61–69, 2006.

[59] H. Holm, F. Ricci, G. Di Martino, E. Bachus, E. D. Nilsson, P. Ballerini, O. Melander, O Hansson, K. Nägga, M. Magnusson, and A. Fedorowski, “Beta-blocker therapy and risk of vascular dementia: A population-based prospective study.” *Vascular pharmacology*, 2020.

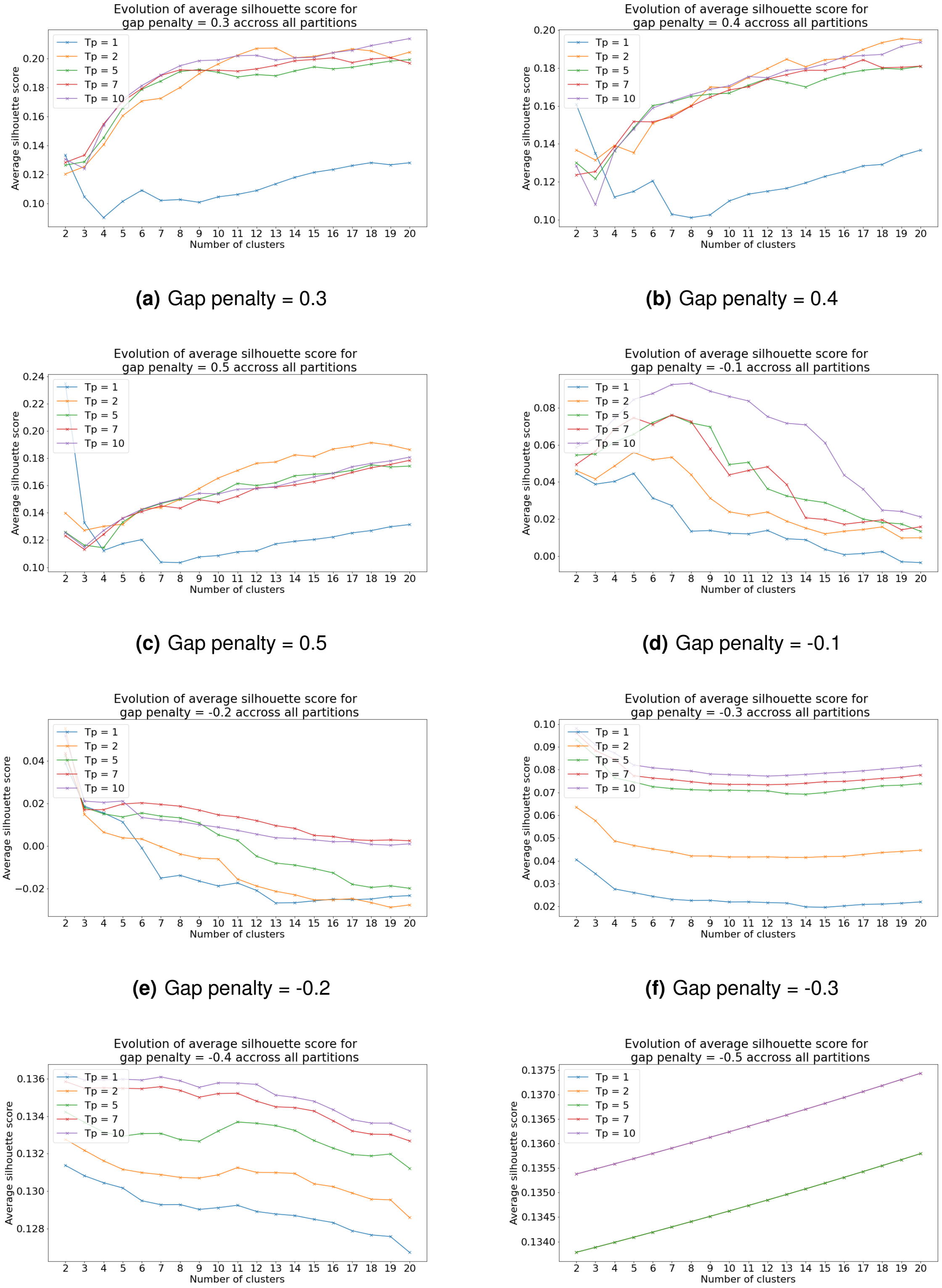
[60] S. Hussain, A. Singh, S. O. Rahman, A. Habib, and A. K. Najmi, “Calcium channel blocker use reduces incident dementia risk in elderly hypertensive patients: A meta-analysis of prospective studies,” *Neuroscience Letters*, vol. 671, pp. 120–127, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0304394018301034

[61] H. Shepherd, G. Livingston, J. Chan, and A. Sommerlad, “Hospitalisation rates and predictors in people with dementia: a systematic review and meta-analysis,” *BMC Medicine*, vol. 17, no. 1, 2019.

A

AliClu parameter validation

Here are presented the remaining plots of silhouette score evolution with the growing number of clusters, for each of the considered gap penalties.



(a) Gap penalty = 0.3

(b) Gap penalty = 0.4

(c) Gap penalty = 0.5

(d) Gap penalty = -0.1

(e) Gap penalty = -0.2

(f) Gap penalty = -0.3

(g) Gap penalty = -0.4

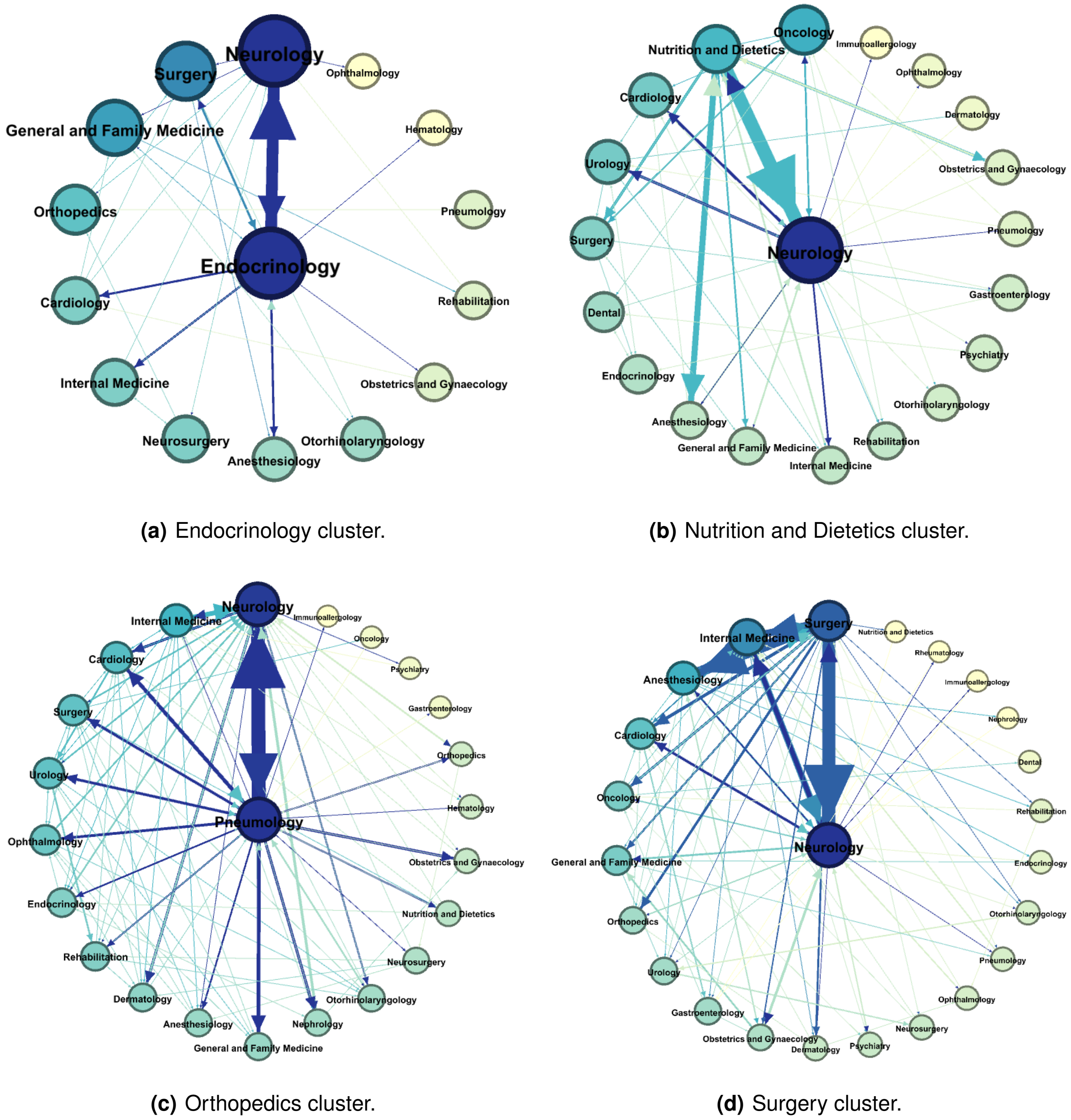
(h) Gap penalty = -0.5

Figure A.1: Evolution of the average SS for each combination of parameters (*g*, *Tp*, *k*), for each gap penalty considered and all values of *Tp* and *k* presented in Table 4.1

B

Directed medical consult graphs and graphs of co-occurrence obtained per cluster

Here we present the remaining directed graphs obtained for each cluster showing the type of consult and transition prevalence, as well as the chronic disease co-occurrence graphs for each subset resulting from AliClu.



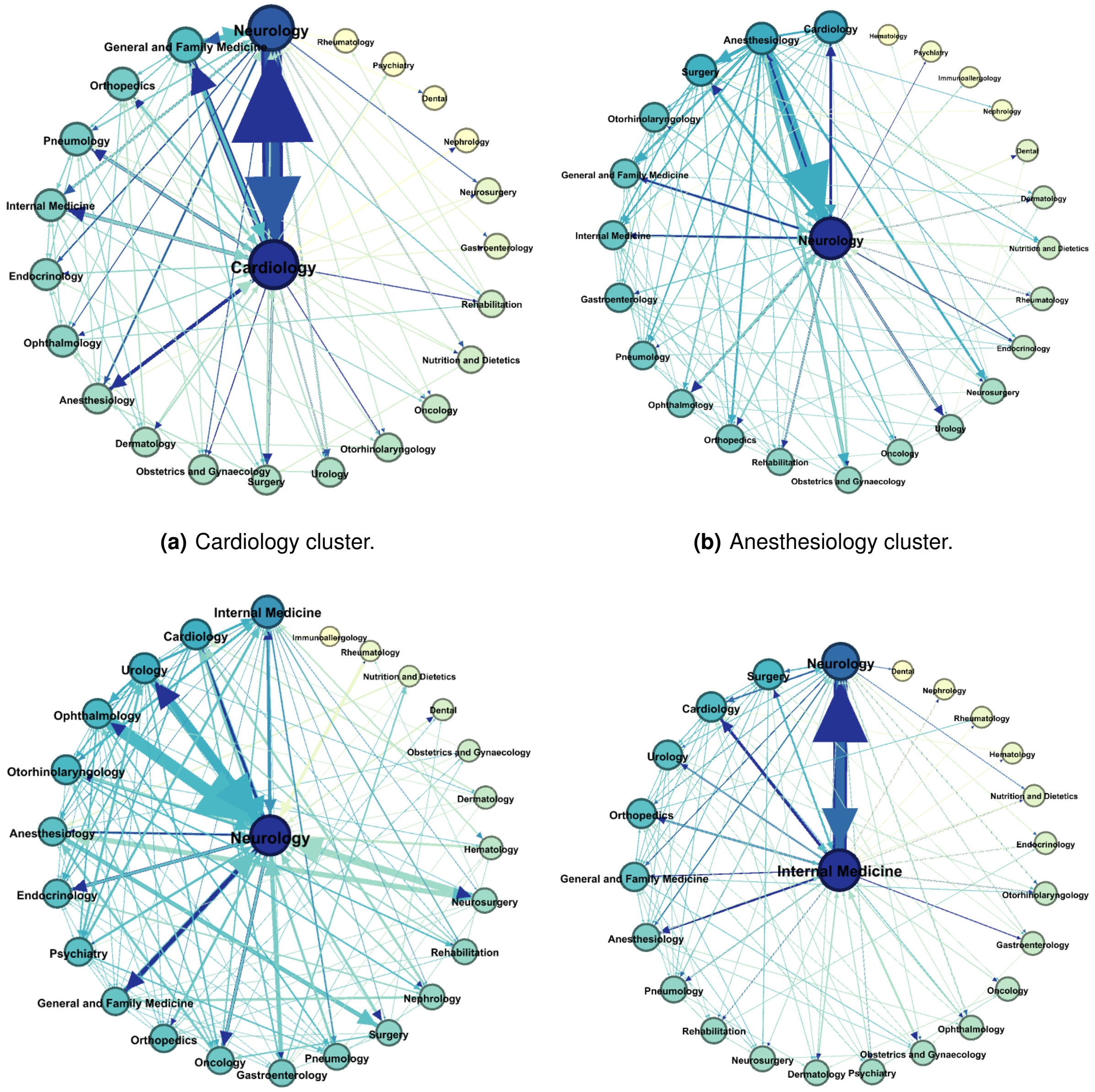
(a) Endocrinology cluster.

(b) Nutrition and Dietetics cluster.

(c) Orthopedics cluster.

(d) Surgery cluster.

Figure B.1: Directed graphs showing the transitions between different medical consults within clusters (a) 1, (b) 2, (c) 4 and (d) 6. Bigger and darker colored nodes indicate higher occurrence of that consult, while wider edges indicate more patients underwent that transition at least once.



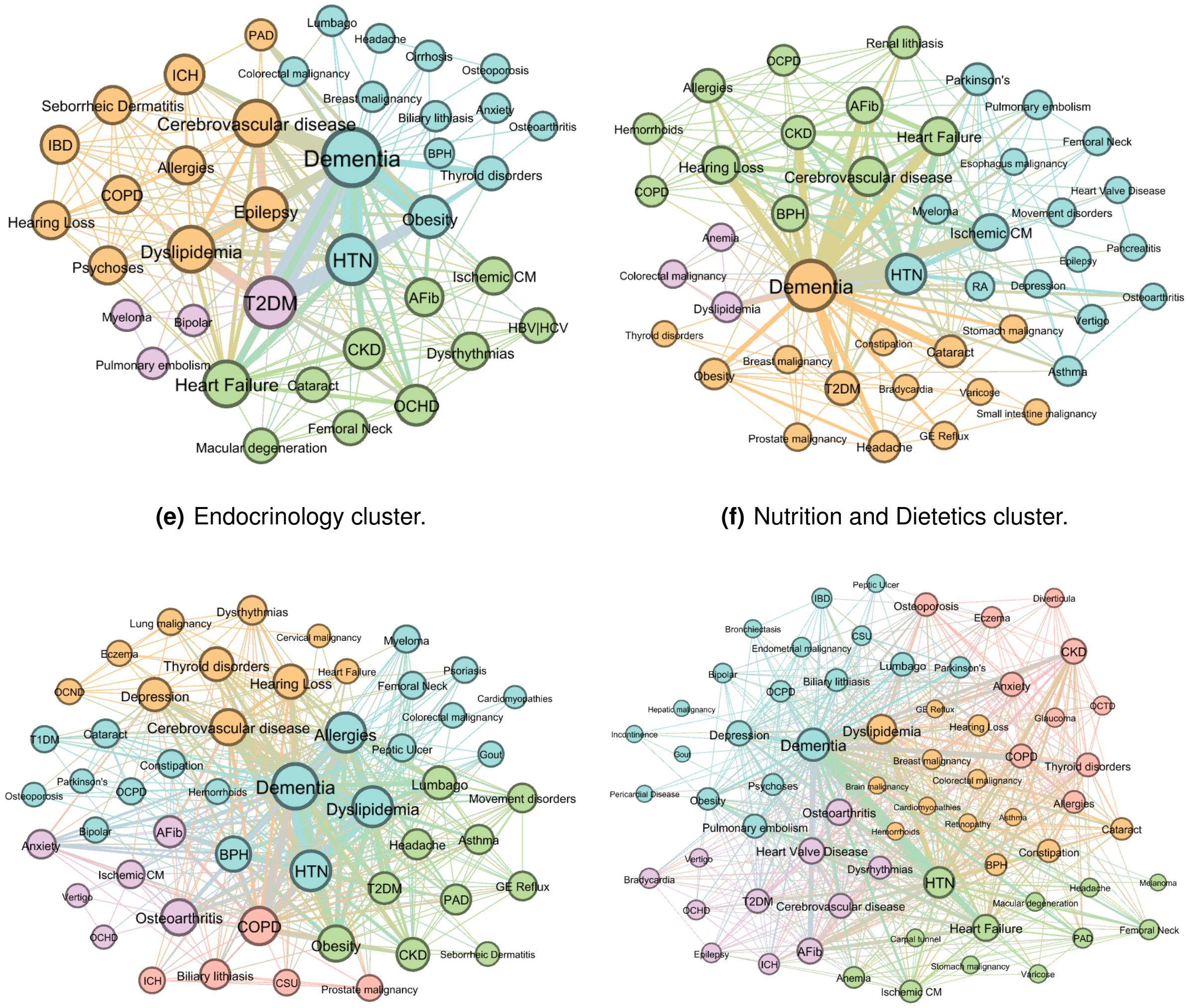
(a) Cardiology cluster.

(b) Anesthesiology cluster.

(c) *Outliers* cluster.

(d) Internal Medicine cluster.

Figure B.1: Directed graphs showing the transitions between different medical consults within clusters (a) 7, (b) 8, (c) 9 and (d) 10. Bigger and darker colored nodes indicate higher occurrence of that consult, while wider edges indicate more patients underwent that transition at least once.



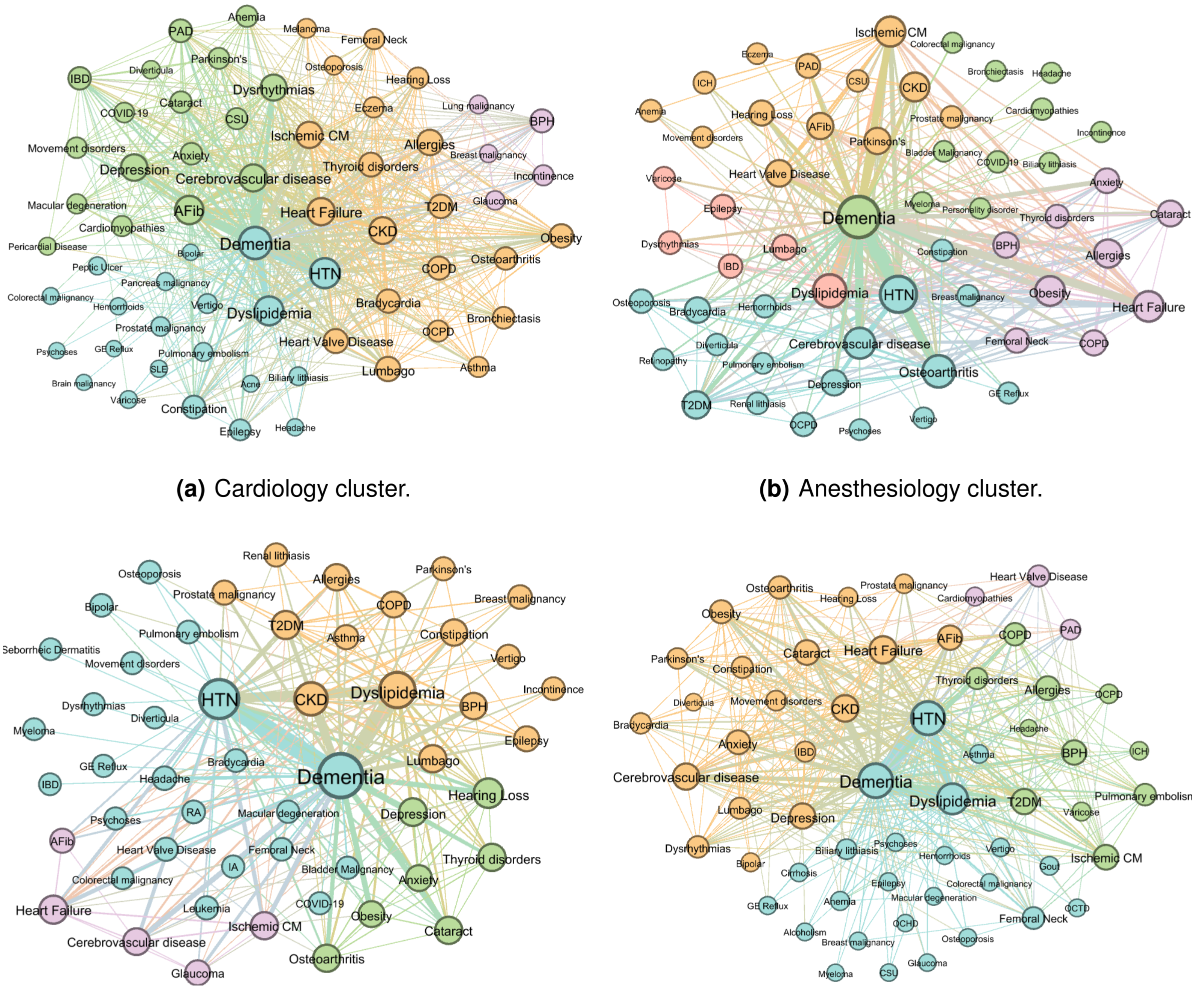
(e) Endocrinology cluster.

(f) Nutrition and Dietetics cluster.

(g) Orthopedics cluster.

(h) Surgery cluster.

Figure B.2: Graphs of chronic disease co-occurrence in patients belonging to clusters (a) 1, (b) 2, (c) 4 and (d) 6. Bigger nodes indicate higher prevalence of that disease in the subset, while wider edges indicate more frequency of co-occurrence of that pair of chronic illnesses.



(a) Cardiology cluster.

(b) Anesthesiology cluster.

(c) *Outliers* cluster.

(d) Internal Medicine cluster.

Figure B.2: Graphs of chronic disease co-occurrence in patients belonging to clusters (a) 7, (b) 8, (c) 9 and (d) 10. Bigger nodes indicate higher prevalence of that disease in the subset, while wider edges indicate more frequency of co-occurrence of that pair of chronic illnesses.