Using Network Science and Clustering for the characterization and stratification of Migraine patients

Maria Manuel Lopes Jacinto

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor(s): Prof. Pedro Tiago Gonçalves Monteiro
Prof. Andreia Sofia Monteiro Teixeira

Examination Committee

Chairperson: Prof. Mário Jorge Costa Gaspar da Silva
Supervisor: Prof. Pedro Tiago Gonçalves Monteiro
Member of the Committee: Doctor Miguel Carvalho Valente Esaguy Coimbra

June 2022

Declaration

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

Acknowledgments

Primeiramente, quero agradecer aos meus orientadores, Prof. Pedro Monteiro e Prof. Sofia Teixeira, por todo o apoio e motivação que me incutiram durante todos estes meses. A ajuda que me deram tornou o trabalho mais fácil e foi essencial para aprender ferramentas que levarei para o meu futuro. Gostaria também de agradecer à Dr. Raquel Gouveia, pela disponibilidade e por todo o feedback que foi essencial para compreender melhor a condição da enxaqueca. Agradeço também ao Miguel Froes e à Catarina Martins por toda a ajuda e tempo que me disponibilizaram.

Num tom mais pessoal, gostaria de agradecer àqueles que me são mais queridos e que sem eles não seria possível nada disto. Em primeiro lugar, quero agradecer à Maria do 2º ano da faculdade, que nunca imaginou chegar a ser Mestre. Com todas as dificuldades e dias difíceis durante todos estes anos, obrigada por não teres desistido e por teres conseguido chegar até ao fim. Que bom que é poder provar que quando acreditamos em nós próprios, somos capazes de tudo. E que bom que é viver.

Os meus amigos foram um pilar essencial nesta jornada, e terei sempre em mente todo o apoio que recebi. Agradeço ao meu grande amigo Jorge, pelas horas e paciência infinita que teve comigo nestes últimos meses. Talvez esta tese não estivesse hoje feita sem o teu apoio e por isso te agradeço muito. Queria também agradecer às minhas queridas Marias, Maria Luísa e Maria Teresa: obrigada por me ajudarem a crescer e acreditarem sempre em mim, especialmente quando eu não o fazia. A amizade que criámos e os momentos que tivemos juntas foram das melhores coisas que a faculdade me deu e que certamente irei recordar durante muitos anos. Aos amigos que fiz e que tornaram estes anos de faculdade menos dolorosos e mais especiais, com carinho particular para a Mary, Pipa, Cat e Margas, um obrigada por todas as memórias: churrascos, bolos, noites de jogos e viagens. Que continuemos a cultivar as nossas amizades mesmo quando a vida acontecer e deixarmos de ter tempo. Agradecer também à Carlos Mardel e à cidade que se tornou a minha favorita e que é Lisboa, onde foi possível ver nascer a versão da Maria que é feliz.

Por fim, e o mais importante de tudo, agradeço a toda a minha família pelo apoio e carinho incondicional durante todos estes anos, e por terem feito e continuarem a fazer tudo o que podem para me dar o melhor possível. Muito obrigada, são tudo para mim. Quero agradecer em especial à minha mãe, que é uma inspiração na minha vida; ao meu pai, que consegue ser sempre a minha luz; ao Pedro e à Rita, por terem o privilégio de serem meus irmãos; aos meus avós, Manuel Jacinto e Fernanda Jacinto, que para além de serem as minhas referências em tudo o que fazem, são as melhores pessoas que vou ter na vida. E por fim, à minha querida avó Albertina, que adoraria saber que a sua neta Marilas é Mestre. Obrigada a todos por acreditarem nas minhas capacidades e me proporcionarem o privilégio de estar onde estou hoje. Talvez nunca vos consiga recompensar, mas fica a imensa gratidão.

O futuro vai ser o que eu quiser.

Resumo

O objetivo desta dissertação é caracterizar e estratificar os doentes que apresentam enxaqueca e compreender de que forma esta condição se relaciona com as suas comorbidades mais comuns, e como se distinguem os pacientes. Utilizando Registos Clínicos Eletrónicos, o trabalho apresentado analisa dois conjuntos de dados distintos: Medical Information Mart for Intensive Care IV, que contém informações sobre pacientes em hospitais dos Estados Unidos da América e e eICU Collaborative Research Database, relacionado apenas com unidades de cuidados intensivos em todo o país. Para avaliar as relações entre as comorbidades mais comuns da enxaqueca, foram geradas redes que conectam estas condições tendo em conta a sua co-ocorrência na população. Com o intuito de agrupar os pacientes que apresentam enxaqueca, foi feita uma analise de clustering utilizando dados demográficos e comorbididades. Com os resultados destas análises, foi possível confirmar algumas diferenças de género associadas a este tipo de pacientes e que constam na literatura, confirmando também a sua complexidade. As redes permitiram extrair as associações mais fortemente relacionadas com enxaqueca que são distúrbio de ansiedade, refluxo gastroesofágico, assim como diabetes e obesidade. As mulheres têm um espectro mais amplo de combinações de comorbidades em relação ao que é visto nos homens. Foi possível também identificar quatro diferentes grupos de doentes, em que um destes grupos manifesta caraterísticas descritas na literatura, onde a idade reprodutiva das mulheres são aspectos-chave importantes; e outro cluster diretamente relacionado com pacientes com multimorbilidade.

Palavras-chave: Enxaqueca, comorbidades, registos clínicos eletrónicos, fenotipagem, redes, clustering.

Abstract

The aim of this dissertation is to characterize and stratify patients with Migraine, understand how this condition is related to its most common comorbidities and how patients can be distinguished. Using Electronic Health Records, the presented work analyzes two distinct datasets: Medical Information Mart for Intensive Care IV, which contains information about patients in hospitals across the United States of America and eICU Collaborative Research Database, related to only intensive care units. To assess the relationships between the most common Migraine comorbidities, networks were generated by connecting these comorbid conditions, taking into account their co-occurrence among patients. In order to group patients with this condition, a clustering analysis was performed using demographic data and comorbidities. With the results of these analyses, it was possible to confirm some gender differences associated with this type of patients which are included in the literature, and also confirm their complexity. The networks allowed us to extract the associations most strongly related to Migraine, which are anxiety disorder, gastroesophageal reflux disease, as well as some other conditions such as diabetes and obesity. Women have a wider spectrum of comorbidities than what is seen in men. It was also possible to identify four different groups of patients, in which one of these groups manifests characteristics described in the literature, where women's childbearing ages are important key aspects; and another cluster is directly related to patients with multimorbidity.

Keywords: Migraine, comorbid conditions, electronic health records, phenotyping, network science, clustering.

Contents

List of Tables

List of Figures

List of Acronyms

APACHE Acute Physiology and Chronic Health Evaluation

EHR Electronic Health Records

eICU-CRD eICU Collaborative Research Database

EMR Electronic Medical Records

HIPAA Health Insurance Portability and Accountability Act

ICD International Classification of Diseases

ICD-10 International Classification of Diseases 10th Revision

ICD-11 International Classification of Diseases 11th Revision

IT Information Technology

MIMIC-III Medical Information Mart for Intensive Care III

MIMIC-IV Medical Information Mart for Intensive Care IV

NPL Natural Language Processing

PCA Principal Component Analysis

PDN Phenotypic Disease Networks

SS Silhouette Score

WHO World Health Organization

YLD Years Lost due to Disability

Chapter 1

Introduction

1.1 Motivation

Women and men can display similarities in health. However, more often than not, there are some significant differences that can be found regarding gender. When studying the population as a whole, without desegregating gender, it may lead to missing key aspects of each gendered population. According to Regitz-Zagrosek 2012, in most health studies, gender is considered as a variable, rather than a striking element of focus. This can be misleading, as the differences between female and male population will go under-looked and may reflect differences in treatment, prevention and management of diseases. Knowledge gaps in gender differences can be determined by a variety of factors and identifying and understanding them can be of great benefit to the population. In Short et al. 2013, it is explained that studying the gender differences that often exist in health may unveil interesting and sometimes vital information for health-based processes and circumstances.

Migraine is a condition that affects the global population and is considered one of the top ten most disabling conditions, as stated in Vos et al. 2017. This challenging condition carries a burden on healthcare and can lead to poor quality of life for the individuals affected by it. There are already several studies which emphasize this condition's burden in the population, from day-to-day lifestyle to work-related implications (Leonardi and Raggi 2019). Although it affects the population at a global scale, Migraine affects women in a bigger percentage and adds a higher severity and long-lasting effects in most cases for the female population (Pavlovic et al. 2017). In fact, women are two to three times more likely to be affected by Migraine (Vetvik and MacGregor 2017). For young women under the ages of 50, Migraine is considered the first most disabling disorder, causing the highest value of Years Lost due to Disability (YLD) according to Vos et al. 2017. There are also some gender-known differences based on age at which this condition appears within individuals, which has been found in Peterlin et al. 2011, for patients who present Migraine. Women's pre and post menopausal stages of life have different prevalences of this disorder, when compared to men. In Berg et al. 2015, which comprises a compilation of possible gender gap knowledge and how to tackle the most relevant conditions and which can be of benefit to the female population, it was possible to assess that Migraine is a condition that needs more exploration. The gender differences associated with this condition were the motivation to deepen the study of migraine and characterize its patients.

Life expectancy has been increasing throughout the last century, and will continue to do so due to advances in Medicine and quality of life stated by the World Health Organization 2019. Alongside a longer life time, the number of illnesses one individual possesses is expected to rise and continue to affect people as the population ages. Multimorbidity is a concept that has a wide range of definitions but can be defined as the co-occurrence of multiple conditions in an individual, as seen by Valderas et al. 2009. The co-occurrence of multiple diseases in patients is not uncommon and should be studied, as it sets back the quality of life of individuals and carries an enormous burden on health care services (Chen et al. 2020). Seeing multimorbidity through a holistic lens, it can be explained not as the sum of all the individual diseases each patient has, but as the combination of those diseases as a whole, as seen in Sturmberg et al. 2021. Thus, rather than focusing on each individual disease, it is given a broader and less limited view to the impact of the combination of the diseases on people. The concept of comorbidity was firstly defined as the occurrence of multiple disorders in relation to an index disease (Feinstein 1970). Comorbid conditions can be paired together within individuals and turn the concept into multimorbidity, meaning that patients that present with more than one comorbid condition can be seen as multimorbidity patients.

Usually, Migraine patients are accompanied by a number of different conditions since Migraine does not often appear on its own, as verified in Altamura et al. 2021. By identifying the most common comorbidities among Migraine patients and relating the interactions between them, we can bring some important knowledge to this population. To study the impact of a specific disease in combination with its most prevalent comorbidities, identifying and grouping together patients with common conditions can be beneficial, along with characterizing patients. The study of the co-occurence of these disorders and the associations between each other can be seen through Phenotypic Disease Networks (PDN) (Hidalgo et al. 2009). This allows for the unveiling of not-so-obvious links between conditions, shown by Chmiel et al. 2014 and Kim et al. 2016. On a patient focused approach, clustering methods can be useful to group together individuals with similar characteristics and understand the population of a specific disease. For Migraine patients, clustering patients has been done in Woldeamanuel et al. 2020 to understand the phenotypes of the possible subgroups within this type of patients. Clustering for comorbid conditions among these patients has been seen through Pellicer-Valero et al. 2020.

In order to firstly study the association of comorbid conditions among Migraine patients, data about patients must be retrieved from health care facilities. Hospital data has enormous potential in order to be analyzed and provide optimization of certain services, as well as tackling complex patients and a wide range of different conditions (Dash et al. 2019). The benefits for hospitals and its users, with respect to big data in healthcare, are usually related to medical records. Now called Electronic Health Records (EHR), this type of structured data comprises information about each patient related to lab results, diagnosis, demographic data, among any other relevant summary or reports about the individual. In this dissertation, there are two different datasets: Medical Information Mart for Intensive Care IV (MIMIC-IV) (Johnson et al. 2021) and eICU Collaborative Research Database (eICU-CRD) (Pollard et al. 2018).

These datasets hold information about patients from hospitals across the United States of America. The eICU-CRD contains data about patients in critical care units, while the MIMIC-IV dataset comprises hospitals' wide EHR about all patients, including critical care units. Exploring these two datasets gives an insight on their population and selecting Migraine patients allows to understand this condition at a deeper level.

1.2 Objectives and Contribuitions

The goal of this dissertation is to select patients that experience the condition of Migraine and their most common comorbidities within the available datasets (MIMIC-IV and eICU-CRD), and characterize the population in order to understand how this specific condition is related to other comorbid disorders. It aims to see how these patients can be grouped together through common similarities and characteristics, for the sake of gathering essential information about them, so as to be able to have a more personalized observation about these types of patients. This analysis can be of great importance, verifying some already known gender differences by incorporating a gender lens, resorting to EHR. Although gender and sex comprise different meanings, the terms will be used interchangeably throughout this dissertation.

The characterisation of the MIMIC-IV population, as well as the eICU-CRD population is presented, giving insight to the type of patients contained in these real-life EHRs. After the selection of patients who present the Migraine condition and related comorbid conditions, the characterization of said patients is performed. Resorting to network science and clustering analysis, it was found the most prominent connections between comorbid conditions of Migraine in Migraine subgroups for both datasets. The presented data and machine learning analysis were performed with the aid of *Python*'s programming language and its libraries such as pandas[1], scikit-learn[2], seaborn[3], networkX[4] and the software Gephi[5] for graph visualization. Gender differences that are described in the literature were seen in the results of these analyses, while confirming that the patients who are associated with Migraine are complex ones.

1.3 Thesis Outline

In Chapter 2, the background of the subjects that are addressed throughout the thesis is introduced. Firstly, introducing the concept of comorbidity and multimorbidity and their impacts on understanding diseases and the interactions between conditions. Focusing on the Migraine condition, describing how it affects the population and some already known comorbid conditions and information about them. Discussing EHR phenotyping, network science and clustering analysis methods, while focusing on the Migraine patients.

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

[2] https://scikit-learn.org/

[3] https://seaborn.pydata.org/

[4] https://networkx.org/

[5] https://gephi.org/

Chapter 3 presents the characterisation of the two available datasets of patients for the general population and a brief explanation on how the data was pre-processed in order to proceed to the analysis.

In Chapter 4, results are shown and discussed. Firstly, the focus is on population of Migraine patients for both datasets. As MIMIC-IV contains a broader type of information about patients, the analysis was done based on this dataset. However, it is also possible to assess a simpler analysis of eICU-CRD at the end of this Chapter 4.4, The exploratory analysis used graph theory and clustering in order to better understand the Migraine subgroup.

In Chapter 5, it is presented the conclusions and future work, alongside with some of the limitations of this work.

Chapter 2

Background

Exploring a specific group of patients that present with the same condition can be of great complexity. In order to aid in this analysis, phenotyping these patients resorting to EHR can facilitate this task and give insight to the patients and relevant common characteristics and otherwise under-looked associations. This chapter aims to explain some of the essential concepts that are explored throughout this dissertation, supporting the information with references. Starting with an overview of the concept of comorbidities, and how understanding their relationships can be beneficial, as well as multimorbidity. Explaining the migraine condition in specific, followed by an introduction to EHR phenotyping, and previous work related to it. Concepts of network science are also explored, as a tool to aid in this analysis, as well as some clustering theory in order to understand how the population can be sub-grouped. Studies relevant to each of the sections are referenced and explored.

2.1 Comorbidity and Multimorbidity

The term comorbidity has ambiguous definitions and is often used differently depending on the context it is inserted into. Although it can have a broader meaning, the concept of comorbidity has been defined by Feinstein 1970 as the presence of more than one distinct health condition in a subject, in addition to an index disease. Patients with a certain condition may have several other conditions that coexist with this index condition. This definition may take into account the order at which the appearance of these conditions might affect subjects in their lifetime. This means that some disorders may appear as one-time-only conditions and others affect patients' throughout their whole life. However, it is a known fact that these coexisting medical disorders in relation to an index disease carry a burden on patients who experience them, affecting how patients live and make use of the health care facilities. With a higher number of associated diseases, there is a higher complexity and thus a frequent use of health institutions. Performing a comorbidity analysis brings out the relationships between these medical conditions, and it is possible to eliminate random occurrences of these diseases or, on the other hand, unveil associations that are important and often overlooked.

The comorbidity concept is highly correlated to multimorbidity, that can be defined as the occurrence of multiple chronic or acute diseases in one subject. Differently to what is defined for comorbidity, this definition does not point out the need for a specific index disease. This definition found in Feinstein 1970, unlike what can happen for comorbidity, time of appearance for these disorders is not taken into account, and the diseases do not have to be linked to one index medical condition. However, these two concepts are associated with each other. As comorbidity takes into account a condition's occurrence to an index one, multimorbidity can be defined as the totality of these combination of pairs for the comorbid conditions. As patients get more and more complex with a higher number of diseases associated to them, this takes up another challenge for health practitioners. Rather than analyzing patients and their conditions as isolated, combining it with co-occurring conditions in a patient and deepening the study of how they are related to each other can be beneficial. The complexity of said task adds to a more difficult assessment of patients, and tackling this concept is of most importance. How conditions co-occur in a patient can have an impact on how individuals have to be assisted, greatly impacting costs of health for hospitals and facilities, as seen through Harrison et al. 2021.

According to World Health Organization 2019, life conditions have improved drastically throughout the last decades, allowing for the life expectancy to rise. Alongside with that, it has become more common for people to have a combination of multiple disorders.

In Valderas et al. 2009, for both the comorbidity and the multimorbidity concepts, there is a definition of morbidity burden, which is related to the effect that these diseases and the co-occurrence of them partake in individuals' lives. This burden severely affects subjects who experience them and is associated with gender, age and health-related aspects. Another important aspect that can be seen in addition to the morbidity burden, is the influence that other non-health related characteristics can affect patients lives, meaning that the environment in which individuals are inserted into, such as socioeconomic, cultural and even patients' behaviors, greatly impacts how complex a patient can be.

2.2 Migraine

Taking into account the knowledge gaps that exist in health regarding gender, a more in depth study of different conditions should be done. The migraine condition shows several factors that can be differentiated for both women and men. According to Vetvik and MacGregor 2017, this condition affects women two-to-three times more than men, with effects lasting a longer time and often times in a more severe way, it is of relevance to understand it. This higher prevalence for women who experience migraine in comparison to men, is believed to be linked to sex hormones and women's most fertile period of time in life, as brought up in Peterlin et al. 2011. Although the pathophysiology between sex hormones and migraine has yet to be understood fully, it has been an recognized that hormones related to women's first menstruation (menarche), menstruation, pregnancy, and menopause are influences of migraine occurrence, as well as the use of hormonal contraceptives, explained in Sacco et al. 2012. Thus, it makes sense that this disorder affects women and men differently throughout their life time. Understanding and defining the sex-related differences that can be seen in patients that present with this condition is of most importance, and can up-bring important aspects of it.

Migraine stems from a complex set of pathophysiological mechanisms, in which many processes are interconnected and lead to the neuronal dysfunction involved in this condition. Cortical spreading depression is thought to be one of the causes of migraine auras. This central mechanism of migraine is explained as a slowly propagated wave of neuronal and glial depolarisation that is followed by a depression, in which brain activity is suppressed. Although there are many uncertainties about how this mechanism is related to migraine, it is hypothesized that migraine without aura is a product of this mechanism in the cerebellum (Cutrer et al. 2012).

Migraine is a neurovascular disorder in which the majority of symptoms displayed by affected individuals include headache attacks, nausea, vomiting, photo and/or phonophobia, and skin allodynia (skin sensitivity). This episodic disorder is one of the most common among the population and the attacks are recurrent. There are two types of migraine: migraine with aura, and migraine without aura. For both types, migraines present the aforementioned symptoms, lasting from four to seventy two hours. However, patients that display migraine without aura are the most common cases, making up to 75% of cases (Cutrer et al. 2012). Apart from the already mentioned difference in the ratio of women affected by migraine in comparison to men, there are some other known sex differences in clinical features of this condition, one of them being the duration at which migraine attacks last, explained by Tonini 2018. Although the consistency at which the intensity and even the frequency of said attacks have been linked to be higher in women, it has not been reported thoroughly, and thus needs further investigation. Symptoms such as nausea, vomiting, photophobia and phonophobia are reported to be seen more in women than in men. One important aspect of how migraine manifests and how the symptoms present in individuals has to do with age. Reports have shown that men have a steadier representation of symptoms, with no significant changes in their characteristics throughout their lifetime, as opposed to women, who tend to have an increase in the duration and intensity of the attacks after the ages of 30. However, some of these findings have to be considered with apprehension, as there can be some social aspects that affect differently how and how often men and women report their symptoms, caused by underlying gender role expectations. In both Peterlin et al. 2011 and Vetvik and MacGregor 2017's studies, it was pointed out that throughout patients lifetime, migraine has a different prevalence depending on the gender of the individuals it affects. It can be seen through Figure 2.1 how starting from early teenage years, at which women get their menarche and start to monthly experience menstruation, until the late 40s, in which menopause is the key point, women have a higher prevalence throughout these years. Puberty years are where the disproportion of prevalence among women when compared to men starts, and there is an obvious decrease in this predominance due to an improvement once menopause starts. Migraine research in this field is still evaluating how and why these differences happen, but sex hormones have been linked to them (Reddy et al. 2021).

Often connected to a wide variety of other conditions, the migraine condition carries a burden to its patients worldwide population, as seen in Vos et al. 2017. In fact, it has been reported by Al-Hassany et al. 2020 that this neurovascular disorder is placed in the top 10 more debilitating illnesses, affecting an estimation of 1.3 billion people. The Global Burden of Disease has set this condition as the first cause of disability for women under the ages of 50. Migraine associates with a high number of different

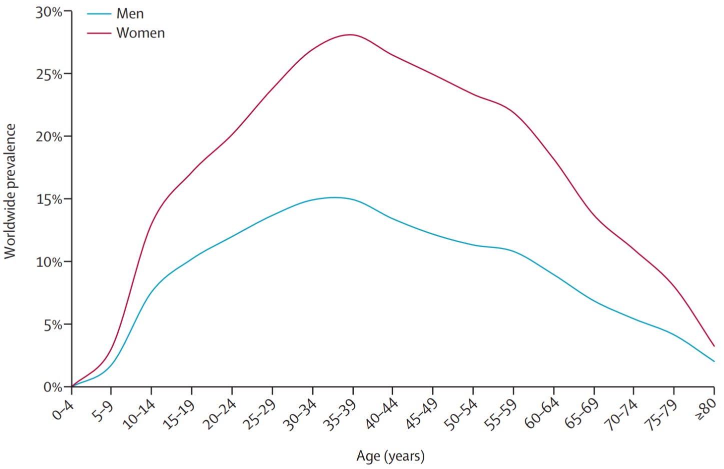


Figure 2.1: Global prevalence of migraine in men and women as reported in Vetvik and MacGregor 2017.

conditions (Steiner et al. 2020). The most common comorbidities of migraine have already been tackled in a review done by Altamura et al. 2021. Seeing how these interact with each other can be of great interest. From gastrointestinal disorders to immunological disorders, to neurological, psychiatric, cardio-cerebrovascular and metaboloendocrine disorders, which are the broader groups of conditions, it can be seen that more often than not, migraine does not present itself on its own in individuals. It is accompanied by a multitude of different disorders and studying how they are related to each other, while taking into account gender, might unveil some interesting concepts that otherwise would have not been explored and identified. The list of the pathologies most commonly associated to migraine has been proposed by Altamura et al. 2021, showing how migraine is connected to these diseases and how they are also connected to migraine in a bi-directional way. A deeper understanding of how they connect to each other can be beneficial as the multiple concurrent comorbidities add complexity to how migraine presents itself clinically and as a prognostic. It is also believed that there is a genetic background for migraine patients, making them more predisposed to this condition and thus to other comorbid conditions related to it, as exposed in Tonini 2018. Some genetic factors may be at the origin of the differences in how migraine affects women differently from men. Non-migraine individuals have a three-times lower probability of not possessing this condition, in relation to people who have relatives that experience migraine (Altamura et al. 2021).

Apart from the correlation to diseases of the central nervous system, as this condition is a neurological disorder, there are other conditions that are of most importance. Patients presenting migraine with aura are considered to be at higher risk of suffering from cardiovascular disorders, especially female patients, as exposed through Kurth et al. 2012. This complex association between migraine and ischemic stroke is linked to mostly patients that present migraine with aura, having no consistent findings in this risk for patients without aura. This is thought to be explained by the mechanism that originates the aura in migraines, which is cortical spreading depression. Migraine is considered to be a higher risk when comparing to other risk conditions such as diabetes, when it comes to experiencing stroke (Schurks et al. 2009). Conditions such as insomnia have also been linked to an increase in prevalence for migraine, found through a study done by Chu et al. 2021. In Amiri et al. 2022's review, some other risk factors have been presented, bringing into awareness that some lifestyle habits can be of importance to patients who experience migraine.

2.3 Electronic Health Records Phenotyping

Information Technology (IT) has allowed for the incorporation of useful tools in healthcare. One of those being Electronic Health Records, which is key for a list of processes in health that can benefit patients, and researchers, as well as practitioners (Dash et al. 2019). Medical records have surged in the beginning of practice medicine and they have evolved to what is now called EHR. Electronic Medical Records (EMR) can be defined as all the information about each patient, from laboratory reports and consults, to current or previous diagnoses, to medical history, medications and treatments plans that occur in a health facility. A major highlight of EHRs differing from EMRs is the fact that EHRs can be used outside of the facility where they were originated, making it an universal record of easy usability (Hayrinen et al. 2008). This data can be generated into two different types: structured and unstructured, the first one being related to coded information such as diagnosis, patients' vital signs, as well as laboratory values; while the latter is related to clinical reports and documents that are written in text which can be clinical notes and more complex documentation (Pendergrass and Crawford 2019).

EHR have surged and have been used in a wide variety of applications, one of them being phenotyping. The concept of phenotyping can fall into many different contexts, but the most common practice is to find cohorts of patients that are associated to a certain disease or a desired characteristic and explore those within said population. This means understanding their phenotypes and clinical features, such as age, gender and ethnicity of patients; how the condition can be related to others, and evaluate risk factors, treatment response of patients, monitoring disease progression, or any other variable that can be of interest. The applications that have been defined through Banda et al. 2018, have fallen into three categories: cross-sectional electronic phenotyping, association phenotyping, and experimental phenotyping. The first one is related mostly to epidemiological research, quality measurements; the second one can be related to the already mentioned evaluation of risk factors, and case-control/cohort studies; and lastly, the experimental type is associated to determining groups of patients eligible for clinical trials.

Phenotyping using EHR, although a useful practice, is also linked to a variety of challenges and blockages in order to archive it. One of the most common reasons that act as a barrier for phenotyping patients, is the difficulty in getting precise EHR among health institutions (Menachemi and Collum 2011). In order for this to be possible, all the individuals involved in the process of phenotyping patients have to cooperate in order to have the most cohesive information about patients among all the contributors. How patients' phenotyping is done by health practitioners is highly related to how accurate the stored data will be, as different health institutions have different methods to create EHR, and thus, is one of the most limiting barriers of EHR.

Since EHR can contain mixed type data, meaning that there can be structured or unstructured data, this adds complexity to the tasks of phenotyping. Unstructured data presents as challenging due to containing abbreviations and a higher chance of typos, since it is written language. There is a need to have a standard phenotyping technique, using machine learning and Natural Language Processing (NPL) based techniques in order to better identity cohorts of patients, as pointed out in Shivade et al. 2014's review. Creating models that can facilitate this task and present more accurate results is beneficial and needed.

2.3.1 International Classification of Diseases

Throughout this dissertation, the mentioning of International Classification of Diseases (ICD) diagnostic codes is used. This is related to a mean of standardized codification for patients that has been developed by the World Health Organization (WHO). Since the 18th century, there has been adopted a way in order to normalize health records and statistics related to diseases in all types of care. This terminology allows for an easier assessment of patients' diagnoses or procedures, granting a basis for comparing health records without the burden of being different for each patient or not standardized in different facilities (WHO 2022).

There are several revisions of ICD, as they are continuously updated, as it is a system that calls for revision, as new conditions, procedures, and even diseases are always being found or old terms no longer being relevant. In order to keep up with these changes, and with the intent of being as simple and easy as possible to analyze and facilitating the accessibility of information between different countries at different times, the ICD system and its different revisions were created. In its essence, this system allows for the comparison and interpretation of data, without compromising its viability and enabling the use of data for health statistics. The most recent and current revision of ICD that has been launched in the 1st of January of 2022 is the International Classification of Diseases 11th Revision (ICD-11). As for this dissertation, older revisions of the ICD classification system were used, namely the 9th and 10th revisions, because they were the only ones available at the beginning of this work and the ones included in the datasets available. In Table 2.1, it is possible to assess the blocks regarding the corresponding chapters that have been divided in order to organize diseases for revision 10 of the ICD code system. Diseases have been grouped together in each chapter and the have a code associated to them that is comprised inside each corresponding block. In Table 2.2, the revision 9 of the ICD system codes and corresponding chapters and titles are shown. Some differences can be pointed out from Tables 2.1 and 2.2, one of them being the number of chapters available, with the 10th revision containing a higher number of chapters. The blocks of codes are also presented differently, where in Table 2.2 the blocks are only digits, and we can see in Table 2.1 that blocks are coded also with alpha characters.

It is important to highlight the differences that comes from both 9th and 10th revisions. This can be assessed in Table 2.3, and it is possible to verify that the 10th revision of the ICD system has advantages in relation to the 9th revision, from the very important specificity of diagnosis, to the number of different codes that exist. While the 9th revision of ICD codes are unable to provide new codes due to the its smaller number of digits, impacting how patients get their diagnosis. With the 10th revision of ICD codes, since it comprises higher number of characters, it is possible to add new codes and includes changes in

Table 2.1: Information about ICD-10-CM diagnostic codes, with corresponding chapters and titles. This can be found at: https://icd.who.int/browse10/2019/en

Blocks	**Chapter and Title**
A00-B99	I *Certain infectious and parasitic diseases*
C00-D48	II *Neoplasms*
D50-D89	III *Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism*
E00-E90	IV *Endocrine, nutritional and metabolic diseases*
F00-F99	V *Mental and behavioural disorders*
G00-G99	VI *Diseases of the nervous system*
H00-H59	VII *Diseases of the eye and adnexa*
H60-H95	VIII *Diseases of the ear and mastoid process*
I00-I99	IX *Diseases of the circulatory system*
J00-J99	X *Diseases of the respiratory system*
K00-K93	XI *Diseases of the digestive system*
L00-L99	XII *Diseases of the skin and subcutaneous tissue*
M00-M99	XIII *Diseases of the musculoskeletal system and connective tissue*
N00-N99	XIV *Diseases of the genitourinary system*
O00-O99	XV *Pregnancy, childbirth and the puerperium*
P00-P96	XVI *Certain conditions originating in the perinatal period*
Q00-Q99	XVII *Congenital malformations, deformations and chromosomal abnormalities*
R00-R99	XVIII *Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified*
S00-T98	XIX *Injury, poisoning and certain other consequences of external causes*
V01-Y98	XX *External causes of morbidity and mortality*
Z00-Z99	XXI *Factors influencing health status and contact with health services*
U00-U85	XXII *Codes for special purposes*

terminology, and combating laterality problems that were found with the 9th revision (Association 2015).

Taking an example of an International Classification of Diseases 10th Revision (ICD-10) code, looking at the condition ”*Migraine*”, its diagnose is linked to the code G43. This falls into the blocks G00-G99, which are the sixth chapter of ”*Diseases of the nervous system*”. If the diagnosis is somewhat more specific than ”G43”, adding more digits to it: ”G43.6”, this adds to the already said diagnostic, meaning that in addition to the Migraine, it is ”*Persistent migraine aura with cerebral infarction*”. As more digits are added, the more specific inside this condition the diagnostic gets. This can be seen in the example:

- G43 *Migraine*
    - G43.6 *Persistent migraine aura with cerebral infarction*
        - G43.60 *Persistent migraine aura with cerebral infarction, not intractable*
            - G43.601 . . . . . . *with status migrainosus*
            - G43.609 . . . . . . *without status migrainosus*

One of the issues associated to the ICD system, is that it can be used as a form of control for billing

Table 2.2: Information about ICD-9-CM diagnostic codes, with corresponding chapters and titles. This can be consulted at: http://icd9.chrisendres.com/

Blocks	**Chapter and Title**
001-139	I *Infectious And Parasitic Diseases*
140-239	II *Neoplasms*
240-279	III *Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders*
280-289	IV *Diseases Of The Blood And Blood-Forming Organs*
290-319	V *Mental disorders*
320-389	VI *Diseases Of The Nervous System And Sense Organs*
390-459	VII *Diseases Of The Circulatory System*
460-519	VIII *Diseases Of The Respiratory System*
520-579	IX *Diseases Of The Digestive System*
580-629	X *Diseases Of The Genitourinary System*
630-679	XI *Complications Of Pregnancy, Childbirth, And The Puerperium*
680-709	XII *Diseases of the skin and subcutaneous tissue*
710-739	XIII *Diseases of the musculoskeletal system and connective tissue*
740-759	XIV *Congenital Anomalies*
760-779	XV *Certain Conditions Originating In The Perinatal Period*
780-799	XVI *Symptoms, Signs, And Ill-Defined Conditions*
800-999	XVII *Injury And Poisoning*
V01-V91	XVIII S*upplementary Classification Of Factors Influencing* *Health Status And Contact With Health Services*
E000-E999	XIX *Supplementary Classification Of External Causes Of Injury And Poisoning*

Table 2.3: Comparison between revision 9 and 10 of the ICD system pointed out by Association 2015.

	ICD-9	**ICD-10**
Length	3-5 characters	3-7 characters
Number of codes	13 000	69 000
First digit	Numeric or alpha (E or V)	Alpha
New codes	Limited addition of new codes	Flexibility to add new codes
Details	Absence of details	Specific
Laterality	Lacks laterality	Possibility to differentiate from left and right

and insurance companies. This can interfere with the representation of some diseases, due to their costs, presenting as one of the limitations for this type of coding system.

2.4 Network Science

Everything in the world is connected. Either through evident links such as family relationships and other not so obvious associations. The most common example is of a social network, represented by people and their connections to each other. Network science approaches said connections from an understanding point of view, trying to grasp more information from more complex concepts that are often too difficult to keep track. The concept of network science is simple, yet it allows for complex analysis of systems, and their dynamics. In this presented section, the main concepts of Network Science are explored, as a basis for the comprehension of this dissertation, resorting to Coscia 2021 for definition of concepts.

2.4.1 Graph Theory

The basis of networks are graphs, which are mathematical representations composed of a collection of nodes - or also vertex, and edges, which are the connections of pairs of nodes. For each graph or network *G*, there is a set of nodes *V* representing entities that can be linked through each other, with the links being represented as the set of edges *E*. The number of nodes is represented by $|V|$, while the number of edges is $|E|$. To put it simply, a graph G is represented by a tuple (V, E), containing a set of nodes and edges that are connected to these nodes. All of this can be refered to as follows: $G = (V, E)$, with $E \subseteq V \times V$. When two nodes i and j are connected to each other through an edge, they are called adjacent or neighbors, and the edge is represented by $e = (u, v) \in E$.

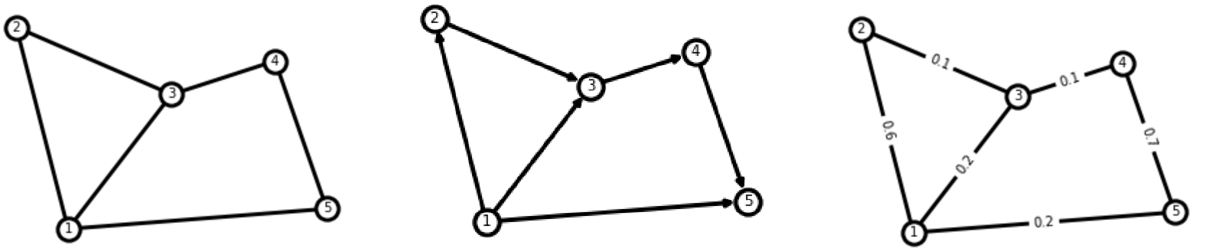
Graphs can have a variety of different characteristics, depending on the complexity or desired outcome of the network. By changing the definition of a simple edge and adding more information to it, other graphs can be made.

When the links between identities are not symmetric, graphs can be directed, meaning that there is only one way for edges. Relationships between nodes that are not reciprocal, can be represented through directed graphs. Thus, the order of connections is important, as it changes the meaning of the network. In this type of graphs, $(i, j) \neq (j, i)$, where nodes are represented as i and j, there is a set of ordered pairs E.

An important concept about graphs is the path, which can be defined as a sequence of connections between nodes. The path represents the connections through edges between node i to node j. The length of such path between the nodes i and j is denoted the distance between said nodes in a graph G, and paths with the shortest length between two nodes i and j are the shortest path. When a path between nodes i and j does not exist, it is denoted as $d(i, j) = \infty$. If the distance between two nodes in a graph G is the longest path, it is assigned as the diameter of the graph. When there is no connection between any nodes contained in graph G, the graph is defined as disconnected. If otherwise, there is at least a path between two any nodes in a graph G, it is a connected graph $G \subseteq V$.

Weighted graphs add another component to the network, assigning a w weight to each edge (i, j), which can represent the distance either proximal or distant. Proximal due to weight aims to group nodes that are closer together and highlight which interactions are significant. For a higher weight, the more unlikely that nodes interact with each other. The relation can be either strong or weak, depending on the weight of the edge. A weighted graph is a tuple represented by $G = (V, E, w)$ and a common example of this type of graphs in real-life application is Google Maps, which finds out directions based on shortest distance of the desired route.

In a graph G, when the elements of E contain binary weight, representing the relations between nodes that could be either negative or positive, it is called a signed network with the weight of an edge e defined as $w(e) \in \{-1, 1\}$.



(a) Simple Graph

(b) Directed Graph

(c) Weighted Graph

Figure 2.2: White circles represent nodes, labeled from 1 to 5 and the black lines represent the edges. (a) represents an **undirected graph**, in which the connections between nodes are reciprocal and do not contain any specific order. As for (b), the arrow indicates in which direction the edges connect to nodes, representing a **directed graph**. Node 1 is connected to node 3, but node 3 is not connected to node 1. In (c), edges contain a weight/measurement associated to between nodes and this type of graph is called **weighted graph**.

When all the nodes are linked to each other, it is called a complete graph $G = (V, E)$. This type of graph is seen in Figure 2.3 and it can also be referred to as *Clique*.

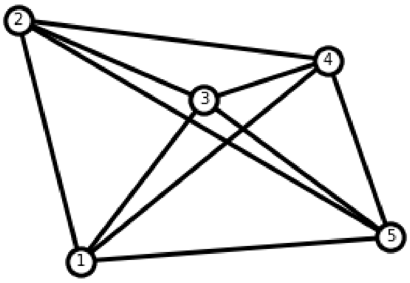


Figure 2.3: A complete graph, containing 5 nodes denoted from 1 to 5.

There are several ways that graphs can be represented. Not only through nodes and edges, but also using matrices. This depends on a variety of factors, such as the size of the graph, the number of nodes and edges, and the desired outcome for the representation of the graph. The most commonly used representations are the adjacency matrix, the incidence matrix, the edge list and the adjacency list.

The Adjacency Matrix represents the graph as a matrix, where each row and column are set to represent a node, and the cells represent the edges. When representing a graph G, the adjacency matrix is a square matrix in which the entries $A_{i,j} \in 0, 1$ obey the following:

$$A_{i,j} = \begin{cases} 0, & \text{if and only if } (i,j) \notin E \\ 1, & \text{if } (i,j) \in E \end{cases} \tag{2.1}$$

The Adjacency List also called Edge list represents the position for a pair of vertexes that are connected, meaning that it is a list of all of the connections between each vertex in a vector of size $|V|$.

Complex networks can too be difficult to analyse. This is where the concept of centrality measures takes part, as it facilitates the analysis of networks properties and allows to take away information from it. Centrality measures can be summarized to how important each node/edge is, inside the corresponding network. As a simple definition, it is an approach that can be used in order to sort through nodes' importance in graphs. There are several ways to determine this importance, basing it on node degree or shortest path, in addition to many others.

The *degree* of a node is an important measure that can be taken in order to find its importance. Usually denoted by *k*, this can be seen as the links and connections each node possesses. The *degree centrality* accounts for the number of edges *k* that a node contains.

The *average degree* accounts for the average node degree of the nodes in a graph, represented by $\langle k \rangle$ and its formula is as follows:

$$\langle k \rangle = N^{-1} \sum_{i=1}^{N} k_i \tag{2.2}$$

Since the degree provides information about each node, combining all the knowledge from all the nodes and randomly choosing one, the probability of said node to have *k* degree is given by the following equation, where N_k represents the number of nodes:

$$p_k = \frac{N_k}{N} \tag{2.3}$$

Closeness centrality is a measure based on shortest path, with an advantage that the closer each node is to other nodes on average, the more central the node is. It is represented by the following equation, where N is the number of nodes and $d(i,j)$ is the length of the shortest path occurring between nodes i and j.

$$C(i) = \frac{N-1}{\sum d(i,j)} \tag{2.4}$$

Modularity is a measure that allows to identify and group together nodes which are more likely to connect to each other. This means that for a graph *G*, there can be several modules/communities in which nodes are distributed into. Nodes inside these communities have a higher modularity when compared to nodes from other communities.

2.4.2 Phenotypic Disease Networks

The concept of PDN was firstly introduced by Hidalgo et al. 2009 and it is a tool that aids in the understanding of phenotype differences inside a demographic population and gives insight about disease progression. In fact, PDN have the ability to unveil links between diseases and comorbid conditions that otherwise could have not been seen, taking a major role in identifying significant relationships between comorbidities. With more information about these relationships, it can bring more benefit to patients. In this type of graph, the nodes usually represent the diseases, while the edges are the connections between said diseases. Depending on the desired outcome of these networks, the weight can derive from different parameters: usually counts of how many times a combination of comorbidities appear, or some other measurement such as Pearson's correlation. PDN have been continuously explored as they provide value to understanding significant relations between the diseases.

In Chmiel et al. 2014, this can be seen as the extinguishing of individual diseases on their own, as this concept no longer makes sense, since diseases are connected to each other and these links can give useful information to patients and clinicians. In this study, multiplex networks were created, which give an insight to the population and their conditions throughout the different stages of life and which are the differences between gender. The links were computed as the probability of having a set of diseases. It was possible to verify which were the most common conditions that affected the population throughout their lifetime, and distinguish it by gender. For this study, basing it on first three digits of the ICD 10th revision codes, patients were divided through three different life phases, depending on their age. It was then possible to observe which conditions were most significant in each of these different intervals of age, taking into account the gender of the population for each of them.

Gender was recognized as a striking differentiation factor and thus studied in Kalgotra et al. 2017. This study aims to identify patients who suffer from multimorbidity and thus have high complexity, and understand how the connections between the co-occurence of these disorders are changed when dividing the population into female and male subgroups. Using network analysis, patients' diagnosis ICD codes from the 9th revision were cut into 3-digitis only, similarly to what was done in Chmiel et al. 2014, simplifying the process of diagnose identification. The results have found some crucial differences between men's and women's multimorbidity patterns. It was possible to verify that women had a higher prevalence of multimorbidity in comparison to men, resorting to networks.

In Jones et al. 2022's review, it reinforces the need to make available more techniques accessible to the public in order to have a standard report for all the possible measures used in multimorbidity networks. In a study done by Kim et al. 2016, networks were built in order to associate diseases of patients using Korean nationwide claims data, using ICD codes from the 10th revision, verifying that there is a need to further analyze these types of networks, in order to understand human diseases and how they're related to each other.

2.5 Clustering

Clustering has been a method used in other domains to uncover subgroups or conditions within a set of elements. This type of unsupervised machine learning method is useful to divide data based on common features (Xu and Wunsch 2010). A cluster depends strongly on the interpretation and where it is inserted, based on an exterior knowledge of this structure, meaning that there needs to be a context in which the cluster is inserted; otherwise it might not make sense. In healthcare, clustering plays an essential role in recognizing subgroups in data that can be seen through a common feature, such as similar patients that comprise the same disease or symptoms. This unsupervised machine learning algorithm is commonly used for electronic phenotyping, incorporating both structured and unstructured EHR data (Ahmad and Khan 2019). By defining how the elements of a cluster are connected to each other through similar grounds, these similarity marks need to be contextualized and understood in order to obtain a true clustering analysis. Depending on the desired outcome, there are several measurements used in order to see how similar elements are, in order to divide them into subgroups. One of the most commonly used proximity measures is Euclidean distance, as to understand how elements of each cluster are connected to each other, but there is also a Manhattan distance measure.

Clustering can be divided into two methods, such as hierarchical and partitional methods.

2.5.1 Hierarchical Clustering

This type of clustering is based on a hierarchy, meaning that a cluster is built using the already pre-formed clusters (Xu and Wunsch 2010). Stemming from the word hierarchy, the purpose of the algorithm is to rank elements together according to importance, in a sequential order. This can have two different approaches, the agglomerative one or bottom-up approach, and the divisive one or top-down approach. For agglomerative clustering, each trait is initially considered an independent single-element cluster. In order to obtain the largest cluster, each of the individual clusters is combined to form the one that is most alike. Thus, elements are merged based on similarity until a final single cluster is formed. This method is the one that is performed more frequently, and comprises a range of different outcomes according to the distance measurements that are used in order to disclose the closest/most similar clusters. In a divisive approach, called divisive hierarchical clustering, it follows the opposite trend, meaning that all elements are initially in one cluster, grouping all elements together, and then for each iteration, it breaks down into the smaller groups. This depends again on the distance between clusters.

2.5.2 Partitional Clustering

In partitional clustering, objects are selected and divided into groups resorting to an initial partition (Xu and Wunsch 2010). This means that throughout sorting, the objects can be changed from one group to another, until a final and optimal group is finished. K-means is the most known method of partitional clustering and opposite to hierarchical clustering, there is no hierarchical component, and elements are grouped together through a series of iterations by calculating and optimizing the position of the center of each cluster, which is called centroid.

2.5.3 Clustering Validation

One of the most important steps in clustering is the evaluation or validation of this method. This too is one of the most challenging aspects associated with this process, as different types of data require different sets of validation and there is not a one-size-fits-all validation for any of the clustering methods. When clustering data, this means that these methods will find groups and associations between the given information, resorting to a specific measure that is given. For this problem, it is needed that a number of desired clusters should be given in order to best portrait the groups found in the data. Identifying the ideal number of *k* clusters is a crucial step and in order to do so, there are some metrics and evaluations that determine which number of *k* clusters is the one that comprises the best results in clustering data, and which reflects on the quality of the information inside each of the formed clusters.

There are three different ways in order to compute the ideal number of *k* clusters, as according to literature: internal, external and relative criteria. As the name suggests, the internal criteria is the one that is used most frequently and it takes into account the data that is being clustered and analyses intrinsic features, whereas the external criteria is related to measurements that have already been known, and the relative criteria can be given through using the same algorithm while inputting different features, or using the same features for different algorithms.

Internal measures have been widely used in order to assess which number of clusters k and which algorithm should be used in order to perform a clustering analysis. A standard practice in order to facilitate this search, is to run the clustering algorithm for different numbers of *k*, and understand in which of these runs it is possible to verify the best internal parameters. The Silhouette Score (SS) is an internal criteria which is widely used in order to understand the number of clusters and its values range from -1 to 1. This is seen in Equation 2.5, where a(i) is defined as the average intracluster distance, measuring the mean distance between i and all elements within the same cluster and b(i) is the average intercluster distance, comprising now the distance between a said element i to all the other points in the closest cluster. By calculating this score, it is possible to assess in which cluster each point should be assigned to, comparing it to others.

$$\text{Silhouette Score } (i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2.5}$$

Some other internal measures are Hubert's statistics, the Davies–Bouldin index, and Dunn's index, all aim to understand the validity of the clusters (Bezdek and Pal 1998).

2.5.4 Clustering for the Migraine condition

Previous work on indentifying subgroups of migraine patients have shown that some clustering algorithms can demonstrate some important features about this type of patients. Other works related to clustering and this condition have been able to identify which are the most common types of headaches among this patients and classify the migraine condition in more depth. Using the already mentioned algorithms, it was possible to find relevant information about some of the most common phenotypes that occur in migraine patients.

In a study done by Tietjen et al. 2007, it took the most common comorbid conditions and analysed within a set population who were identified to have migraine to see which where the connections between these disorders. This study analysed 223 migraine patients who attended a university of headache clinic within the time period of 2 consecutive years, from September 2003 to September 2005. In addition to understanding groups formed around the comorbid conditions, the phenotypic profile of patients includes characteristics related to demographics, psycho-social and headache related characteristics among each of the found groups. In order to identify these migraine constellations of comorbid diseases, a sequential clustering approach was initially used, followed by the traditional agglomerative hierarchical clustering. It was possible to asses three different clustering groups, where one of the groups showed a lack of comorbid relations, while the other two displayed one well-defined disorders and the other with clinically defined disorders. This means that there are some conditions such as hypertension, hyperlipidemia, diabetes mellitus, and hypothyroidism, which can be fully defined and grouped together within these patients. While some other conditions such as fibromyalgia, depression, and anxiety disorder are harder conditions to define because they lack anatomic pathology in tissues and there are no objective findings on exams, and often linked together as found through this cluster. Within these found groups of migraine patients, age played a part, as well as gender, in differing populations.

Although not directly related to migraine, a study was done by Pellicer-Valero et al. 2020, mapping the groups of a total of 208 tension-type headache patients. This allowed to understand features such as headache frequency, duration and intensity, levels of depression and anxiety, health related quality of life and sleep, thresholds on pressure pain, dynamic pressure, headache-related burden among many others. The crucial difference found between the obtained clusters layed on the headache frequency of groups, where one cluster had chronic tension type headache patients, while the two other clusters comprised groups of inviduals who had episodic tension type headache.

A cross-sectional clinical study was perfomed in order to understand the phenotypes that occur naturally among chronic migraine patients, by Woldeamanuel et al. 2020. Identifying 100 patients with chronic migraine, hierarchical agglomerative clustering was performed, as well as a principal component analysis, as to understand natural groups present within this set of patients. For the first analysis, three clusters were formed, where one of these has a high level of exercise and the lowest impact related to migraine, another cluster where depression and migraine-related disabilities rank highest, while the last cluster was a combination of the previously mention two clusters. As for the Principal Component Analysis, it was possible to assess that the features that played the biggest role to explain 65% of the principal components variability were the first 5, while the first principal component was directly related to migraine-related disability features.

On a different context, there were some other studies performed that are not directly related to migraine clustering, but that have been found to be useful for this condition. In Schürks et al. 2011, a Principal Component Analysis (PCA) was performed in order to understand the migraine phenotypes among women and understand the importance of these features in migraine. Another important study for defining the headaches types was done by Diehr et al. 1982 and in order to validate the diagnostic criteria for migraine and tension-type headache, a clustering analysis was done in Bruehl et al. 1999.

An overview of these articles can be seen in Table 2.4, with the title, authors, clustering methods used, features and outcomes.

2.5.5 Clustering mixed-type data

The combination of both continuous and categorical data partake a challenging concept for clustering (Ahmad and Khan 2019). For continuous information such as age, number of disorders, frequency in hospitals' visits, among others, these consist of values and thus can have ambiguous numbers depending on the context they are inserted into. For categorical data, such as gender (Female/Male) or characteristics such as ethnicity (White/African American/Asian,...), these are also a challenge as they do not have a fixed value.

When clustering, it is essential to have an equal contribution among the used features to perform said methods. This means that having continuous data is challenging and there is a need to standardize

Table 2.4: Overview of clustering methods used in migraine-related studies.

Title of paper	Authors	Clustering method	Features	Results
Migraine Comorbidity Constellations	Gretchen E. Tietjen et al. 2007	Sequencial clustering approach, traditional agglomerative hierarchical clustering	Comorbidities	3 different migraine groups were found
Patient Profiling Based on Spectral Clustering for an Enhanced Classification of Patients with Tension-Type Headache	Pellicer-Valero et al. 2020	Spectral clustering	Age, headache related diary, sleep quality, health-related quality of life, pressure pain thresholds (PPT), headache-related burden, etc.	3 clusters identifying different frequencies for headaches for migraine patients
Exploring Natural Clusters of Chronic Migraine Phenotypes: A Cross-Sectional Clinical Study	Woldeamanuel et al. 2020	Hierarchical agglomerative clustering, Principal Component Analysis (PCA)	Demographics (age, gender), lifestyle behaviors, disorders levels	HAC identified 3 clusters, PCA analysis revealed a pattern in clinical features
Migraine features, associated symptoms and triggers: A principal component analysis in the Women's Health Study	Schürks et al. 2011	Principal Component Analysis (PCA)	Demographics, migraine diary (duration, severity, number days), associated symptoms	Features, symptoms, and triggers of migraine are highly correlated
Clustering Analysis to Determine Headache Types	Diehr et al. 1982		Symptoms and frequencies	Two solutions: 2 clusters, 8 clusters
Use of Cluster Analysis to Validate IHS Diagnostic Criteria for Migraine and Tension-Type Headache	Bruehl et al. 1999	Hierarchical agglomerative clustering with Ward's method	Headache symptoms	2 clusters confirming the classification criteria

this information in order to have a balanced data. Categorical data is easier to handle, as they can be dummy-coded and facilitating the process. Where both these types of data are present, Agglomerative Hierarchical Clustering is often used with Gower's distance. This measure is a Manhattan distance measure based one, rather than an Euclidean one. It is possible to compute this distance resorting to Phyton package gower, and for both categorical and numerical data, a value is assigned. Gower's distance has been widely used for clustering methods, and as it is presented through Gower 1971, Gower Similarity measure is defined by the following equation, where *m* refers to the features and *ps* the partial similarities, for i and j elements:

$$GS_{ij} = \frac{1}{m}\sum_{f=1}^{m} ps_{ij}^{(f)} \tag{2.6}$$

As seen through the Ahmad and Khan 2019 review, some studies have been done by applying Gower's distance while using agglomerative hierarchical clustering methods.

Principal Component Analysis

PCA is a method used in order do reduce the dimensionality of large sets of data, while maintaining most of the information (Jolliffe and Cadima 2016). By reducing the number of variables, this accounts for a simpler and easier way to explore data. However, this comes at a cost of a loss in the meaning of the variables that will be grouped together. In order to perform this method, the data must be standardized, as it is crucial for features to have an equal contribution. In this process, features are grouped together and reformulated in order to form the principal components of the dataset. These principal components account for the variability of the data and the initial principal components are the ones that can explain the higher percentage of data of the original features used. As seen in the previous subsection, this method has been widely used in migraine related clustering studies, meaning that it can give an important insight to this condition.

Chapter 3

Dataset Preliminary Analysis

In this Chapter, the characteristics of datasets will be addressed, summarizing the content of each and bringing some light on the type of patients contained in each dataset. Demographics are important to explore, in order to deepen the understanding of the population and take advantage of what is brought through the dataset. After the data exploration for the whole population of both datasets, it was narrowed down to the migraine population, as a crucial step before performing the desired analysis. This specific analysis allows for the understanding of this subgroup and finding some more relevant aspects of what can be seen in Section 3.2.

3.1 Characterization of datasets

MIMIC-IV (Johnson et al. 2021) and eICU-CRD (Pollard et al. 2018) are the two datasets which are analysed and compared throughout this dissertation. Each dataset contains unique information about patients that were admitted to intensive care units throughout a given period of time. Since patients in this type of care are being continuously monitored throughout their hospital stay, the acquisition of the data is simplified and thus, there is a large amount of medical data that can be explored and used in research. When dealing with medical data, there are certain measures that must be taken into account to ensure the protection of patients' rights. For this reason, in order to obtain access to the available databases, a certification was demanded.

One of the main differences between the presented datasets lies in the way the data has been collected. MIMIC-IV's information originates from two different sources, which are an intensive care unit-specific database and a hospital-wide EHR, which can include information from laboratories and data from other types of specific hospital units such as emergency departments. Thus, it does not only contains information about patients in critical care units but inside their wide hospital EHR. However, for the eICU-CRD database, it only refers to patients related to critical care and the data is collected from multiple critical care units across the country. This reflects differences in the results and outcomes related to the analysis of both datasets.

A process that is required and was performed prior to the launching of both datasets was the de-identification of patients. This data anonymization process is the required step in order to assure protection, accordingly to the Health Insurance Portability and Accountability Act (HIPAA). All the available datasets have been previously de-identified, and all compromising information that could potentially lead to the recognition of individuals, such as name, address or telephone number, have been thoroughly removed and no further investigation should be led in order to identify them. Dates and ages have also had some changes in order to solidify this safety.

Table 3.1: Information about datasets, such as a short characterization, the period of time in which the data was collected, number of admissions and patients, and revision of the ICD system codes used for patients' diagnosis.

Dataset	Characterization	Period of time	Number of Admissions	Number of Patients	ICD revision
MIMIC-IV - Medical Information Mart for Intensive Care IV	Updated and more comprehensive revision of the MIMIC-III database, containing critical care information about Intensive Care Units at the Beth Israel Deaconess Medical Center (BIDMC), in the United States of America	January 2008 to December 2019	523 740	382 278	9 and 10
eICU Collaborative Research Database (v2.0)	Database consisting of information about critical care units across the United States	January 2014 to December 2015	200 859	139 367	9

More information about each of the datasets can be found in Table 3.1, such as the period of time in which the data was collected, the number of patients, their admissions in the considered timeline and the revision of ICD codes that can be found in each of them. The period of time in which the data was collected differs for both, as for the MIMIC-IV dataset this interval is larger than in eICU-CRD. However, this does not reflect a major difference in the number of total patients inserted in each of the datasets, nor admissions. As for the ICD revision of these datasets, MIMIC-IV contains both revisions of the ICD system, meaning that a patient can contain diagnoses in both of the revisions. While for eICU-CRD, the dataset codifies for the 9th revision as it is its main way of diagnosing ICD system, but has a translation for the 10th revision within the dataset.

3.1.1 MIMIC-IV

Comprising information about over 320 000 patients in a total amount of 27 tables divided into three different parts, the MIMIC-IV is a database containing details of the Beth Israel Deaconess Medical Center, in the United States of America, containing information about the hospitals' wide EHR, critical care units and emergency departments. This dataset is an updated version of an already released dataset called Medical Information Mart for Intensive Care III (MIMIC-III), adding more information about patients in a structured and easier approach, while dividing clearly the origin of the collection of data. However, the older version of this dataset comprises information about patients who were admitted to critical care units, while the updated version contains information about each of the hospital's electronic health records, in addition to the critical care units, as well as a module dedicated to emergency departments. This means that the structure allows us to understand the source of the information that is presented within patients. There are now five available major modules in the MIMIC-IV database: *Core*, related to the hospital's overall patients' data, such as demographics, admissions and transfers of patients; *hosp*, comprising the hospitals' wide EHR, such as laboratory results, billed diagnosis of patients, and microbiology results, including all the information of patients throughout their stay, from laboratory reports, to medication, procedures; the *ICU* module, which refers to the information about patients in intensive care units; the *ED*, containing the data related to emergency departments patients; and finally the *CXR* module, which related to imaging studies and can be linked to the other modules.

The collection of the medical data was followed from the months of January of 2008 until the month of December 2019, following up to more than three hundred thousand patients' information and a total of more than five hundred thousand admissions. As visible in Table 3.2, this dataset contains information about a total number of 321 406 patients, which contain an equal representation of both men and women, and in which 53 150 are related to the critical care units and 216 467 are related to the emergency departments. The same patient can be admitted at different times, resulting in a higher number of admissions than the total number of patients.

For the MIMIC-IV dataset, in order to comply with Health Insurance Portability and Accountability Act (HIPAA)'s norms, dates have been shifted to years in the future, and grouped into periods of time. Date shifting is a necessary step to ensure the protection of individuals, and for each unique patient, it was a set a random year into the future, bringing consistent information inside the same individual's data. In addition, ages have also been changed, meaning that patients older than 89 years have all been grouped together in the database with the value of age of 91. All these steps ensure that the patients can remain anonymous and the data can still be used in a safe and consistent way.

Table 3.2: Information about number of patients and admissions to either the hospital, critical care units and emergency departments of MIMIC-IV database. hosp - hospital module, ICU - Intensive Care Units, ED - Emergency Departments.

		Number of patients	**Number of admissions**
hosp	Female	170 017	523 740
	Male	151 389	
	Total	321 406	
ICU	Female	23 353	69 211
	Male	29 797	
	Total	53 150	
ED	Female	115 446	213 834
	Male	101 021	
	Total	216 467	

For all the *.csv* tables contained in this dataset, they can easily be connected through a series of different identifiers. There are three major ways to identify a patient in this dataset: through a *subject_id*, a *hadm_id*, and *stay_id*. Each unique patient gets assigned a *subject_id*, which consists of the anonymous

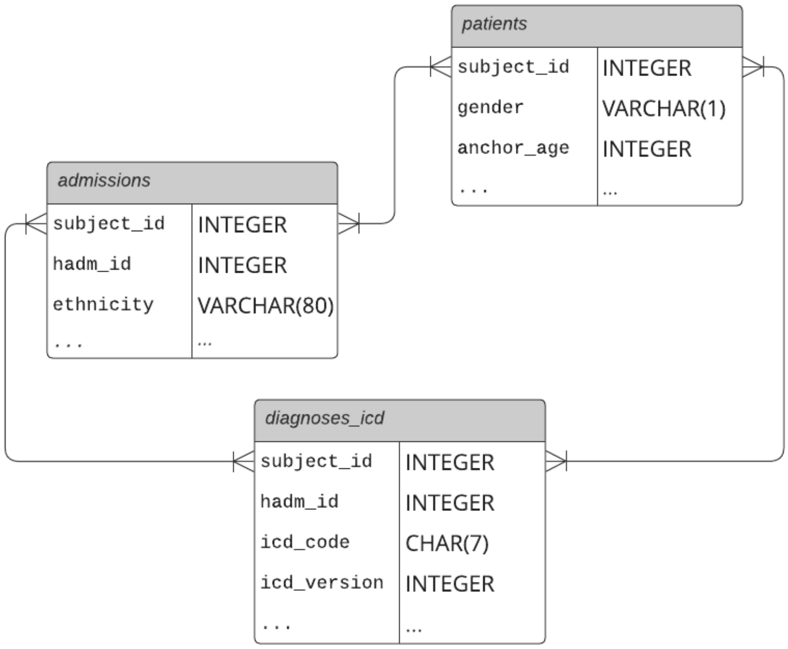
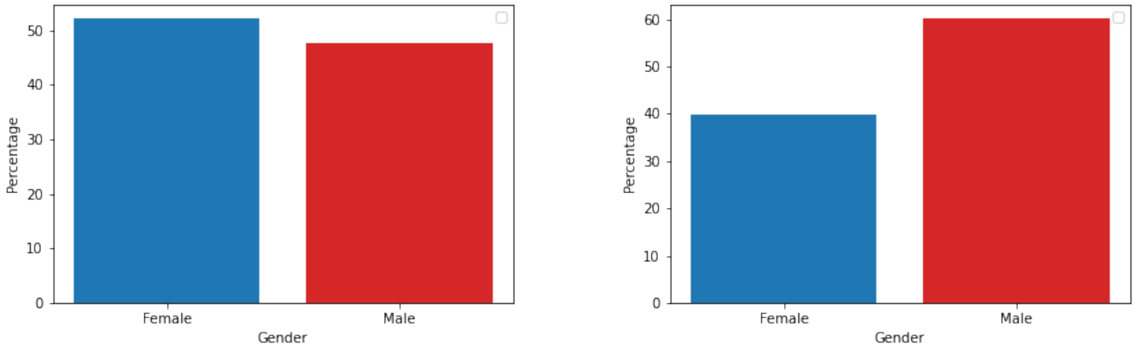


Figure 3.1: Easily accessible information about MIMIC-IV *.csv* tables through: https://mimic.mit.edu/

medical record number for an individual. For the hospitalizations of said patient, it is assigned another important identifier, which is *hadm_id*. Each of these identifiers can be connected to one patient only, resuming to the *subject_id* as the most important identifier. For the practicality of the process, this is the identifier that is selected and used in order to group and analyze the patients in this dataset, and connect to tables with the relevant information for this thesis. For the purpose of this dissertation, only three of the tables were used in order to analyze the patients: admissions, patients and diagnoses. These were tables that are related to the *core* and *hosp* modules. How these tables are connected to each other can be seen in Figure 3.1. As previously mentioned, the identifier used in order to connect all the tables within the MIMIC-IV dataset is the *subject_id*. More information about the tables can be analyzed online and is available to the public. Using only the relevant tables for this analysis, it was possible to get a demographic look at the patients in this dataset.

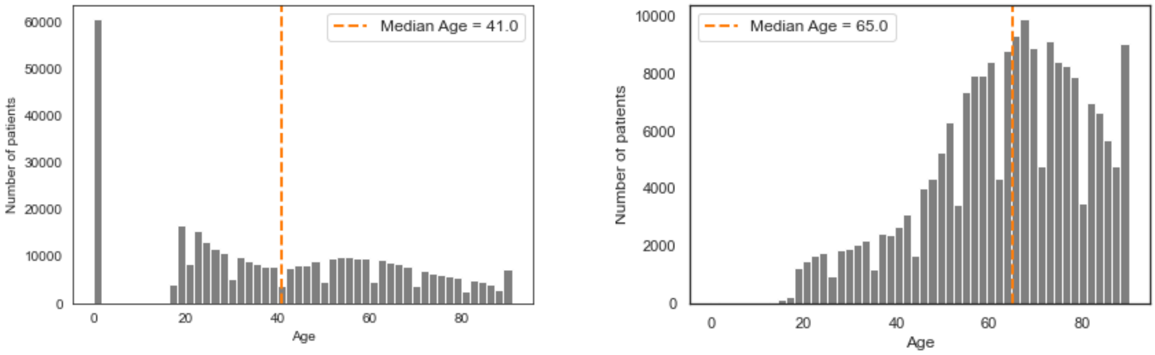
When it comes to the gender of the overall patients, in Figure 3.2 (a), it is seen that the majority of patients are distributed equally for both men and women, confirmed by the already mentioned number in Table 3.2. When analyzing the age of the MIMIC-IV patients, it can be seen in Figure 3.3 (a) that there is a great number of patients with age of 0, and this is related to the newborn babies found in the EHR of the hospital. From the ages of 0 until 18 years old, there is no data regarding age, as MIMIC-IV contains no patients in this dataset who presented said ages. The median age of MIMIC-IV patients is 41 years old. As for the purpose of further analyzing patients, only patients with ages above 18 years old were considered, as MIMIC-IV's information about infants is not relevant to this dissertation and these were disregarded. Information about patients' ethnicity can be found in Table 3.4 (a), where up to 60% of the patients in this database are predominantly white, with the second-highest percentage of patients being Black/African American. For this dataset, there are 9 types of admissions, and the percentage of each of



(a) MIMIC-IV dataset

(b) eICU-CRD dataset

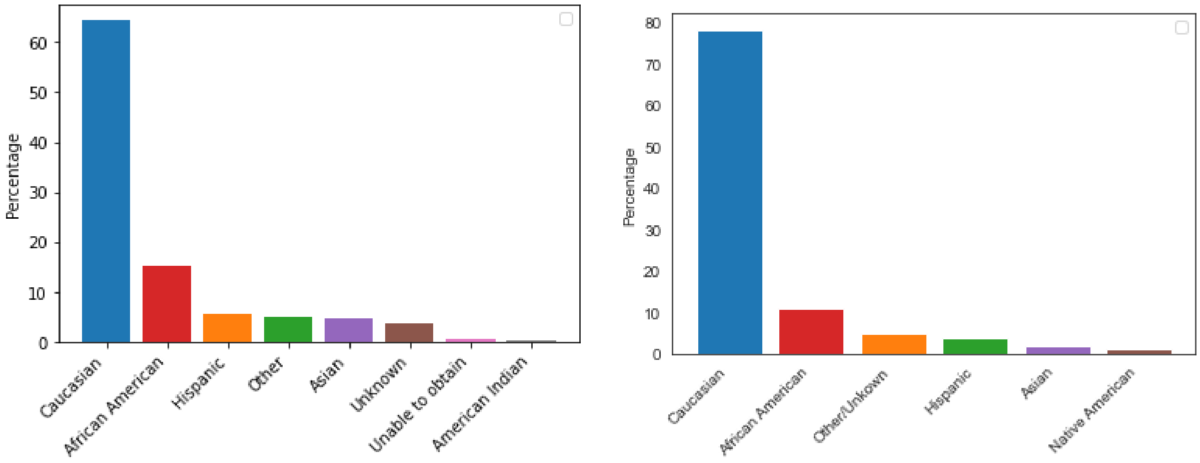
Figure 3.2: Percentages of female and male patients that are in the MIMIC-IV dataset in (a) and in the eICU-CRD dataset in (b).



(a) MIMIC-IV dataset

(b) eICU-CRD dataset

Figure 3.3: Number of of patients that are in the MIMIC-IV dataset in (a) and in the eICU-CRD dataset in (b). The orange line represents the median age of all subjects in each dataset.

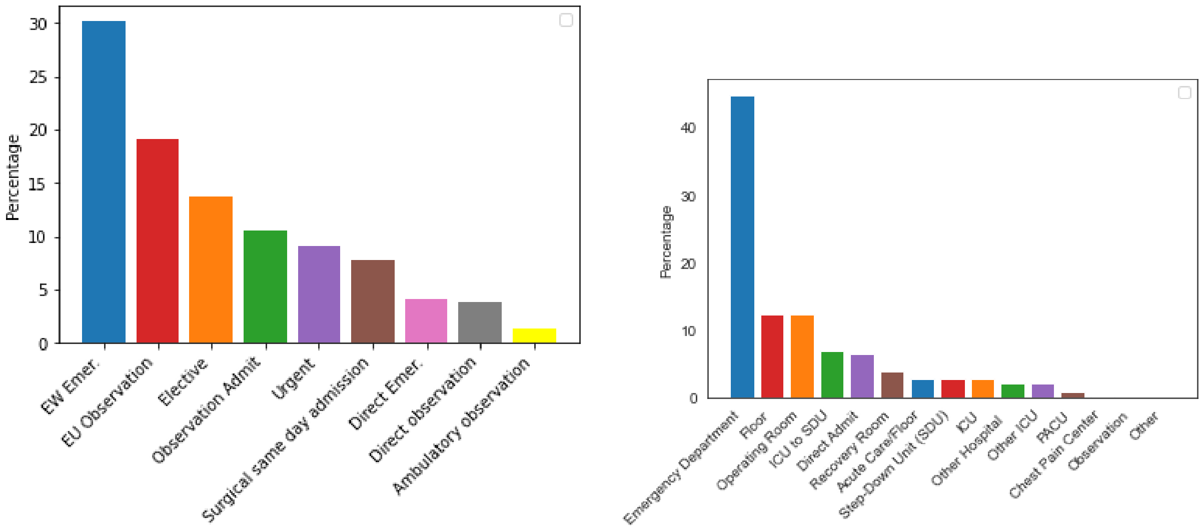


(a) MIMIC-IV dataset

(b) eICU-CRD dataset

Figure 3.4: Percentages of patients that are in the MIMIC-IV dataset in (a) and in the eICU-CRD dataset in (b) based on their ethnicities.

them can be seen in Figure 3.5. The highest number of admissions is through EW, which is Emergency Ward.



(a) MIMIC-IV dataset

(b) eICU-CRD dataset

Figure 3.5: Percentages of patients that are in the MIMIC-IV dataset in (a) and in the eICU-CRD dataset in (b) based on their admission types.

3.1.2 eICU-CRD

The dataset presented in this subsection is called eICU Collaborative Research Database. This contains information about critical care patients who were admitted to these type of units across the United States of America throughout the years of 2014 and 2015. Similar to what happens in the MIMIC-IV database, in order to de-identify patients and their admissions to the hospital, all the data that could potentially lead to the identification of patients has been thoroughly removed, including hospital and units identifiers. The acquisition of data of patients from critical care is facilitated due to the continuous monitoring of patients in these units, leading to the collection of a large amount of data about said patients.

The eICU-CRD data is distributed throughout 31 *.csv* documents, and it contains information about patients' diagnosis, vital sign measurements, what type of care plan each patient is associated with, as well as details about treatment and Acute Physiology and Chronic Health Evaluation (APACHE) measures which is related to the severity of illnesses. Identically to the MIMIC-IV dataset, there are a series of identifies throughout all the tables in the database. The *hospitalid* identifies each hospital in the database, the *uniquepid* codes for each patient, *patienthealthsystemsstayid* for each stay at the hospital, *patientunitstayid* identifies the unit stay. The latter is the one that is found throughout all the *.csv*s and is defined as the primary identifier across the database. Similarly to what is shown in the MIMIC-IV database, we can see through Figure 3.6 how the *.csv* tables are connected to each other. In this case, only the tables related to patients and diagnosis were analyzed. The patients' table (patients.csv) contains all the demographic information about them, and the diagnosis table related to each patient's diagnosis codes.

It is possible to verify a few features of this sample of patients and characterize them. For a total of 139 367 patients, and a total number of 200 859 admissions, it can be verified in Table 3.3 that 92 303 are related to women and 108 379 to men. This goes accordingly to what can be seen in Figure 3.2

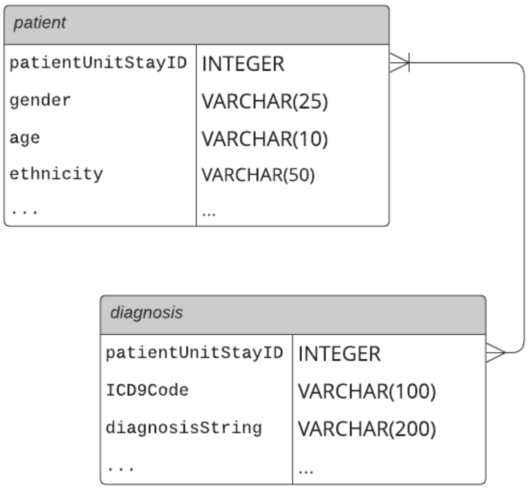


Figure 3.6: Easily accessible information about eICU-CRD *.csv* tables through: https://eicu-crd.mit.edu/about/eicu/

(b), as the distribution shows that men have a higher percentage in this database. As seen in Figure 3.2 (b), men take a higher percentage of the total number of patients in this dataset, in comparison to women. One particularity is that there are two more variables in the gender component of this dataset represented as *Unknown* or *Other*. For the purpose of this dissertation, these were eliminated from the desired data to analyze, as they do not add useful information.

Table 3.3: Information about number of patients and admissions to either the hospital, critical care units and emergency departments of eICU-CRD database.

		Number of admissions	**Number of patients**
eICU-CRD	Female	92 303	139 367
	Male	108 379	
	Total	200 859	

Regarding age, the patients in this dataset have a median age of 65 years, being an older population than in the MIMIC-IV dataset and it is also possible to assess that older ages play a higher role in comparison to MIMIC-IV. This can be seen in Figure 3.3 (b). However, similar to what happens in the MIMIC-IV dataset regarding patients older than 89 years, this is also a measure that has been taken in this dataset. Thus, patients who were older than 89 years old were assigned the age of 90, which is the highest value for age in this dataset and it can be verified that there is a high number of patients in this situation. Similarly, there are no patients with ages below 18 years old.

In Figure 3.4 (b), it is possible to see to which ethnicity each of the patients in this dataset is assigned to. Similarly to MIMIC-IV, the predominance of white people is higher and the percentage, in this case, is up to almost 80% of the patients, followed again by Black/African American people.

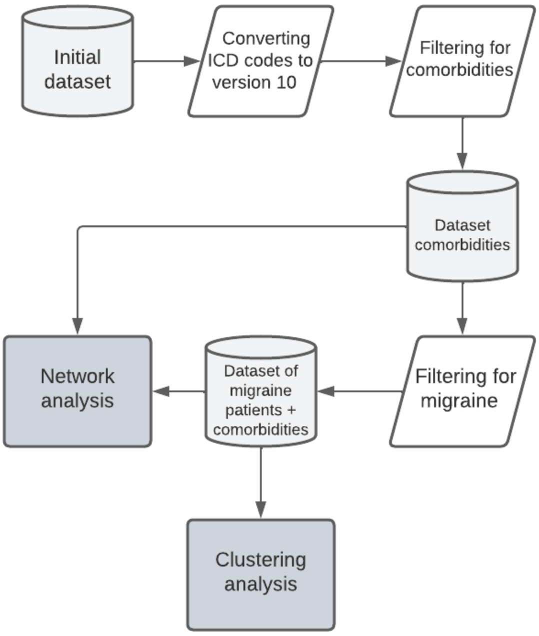

Initial dataset
Converting ICD codes to version 10
Filtering for comorbidities
Dataset comorbidities
Network analysis
Dataset of migraine patients + comorbidities
Filtering for migraine
Clustering analysis

Figure 3.7: Pipeline to obtain the prepared datasets for both network and clustering analyses.

3.2 Pre-processing the Datasets

Prior to deepening the study and performing any type of analysis on both datasets, it was necessary to go through a pre-processing stage. In order to be able to analyze the population of each of the datasets while specifying the migraine condition, there were some pre-processing steps that were taken. In Figure 3.7, it is possible to assess which were the relevant steps in order to obtain all of these analyses. Patients' diagnoses codes in the MIMIC-IV dataset were a mixture of both ICD revisions, meaning that a patient could contain codes of both the 9th and the 10th revisions. While for the eICU-CRD dataset, it contained a diagnosis *string* with the ICD code for both revisions, thus making it possible to skip the conversion step for these patients. In order to have consistent information about patients for both datasets and facilitate the upcoming processes, it was necessary to have the diagnosis codes to only contain one of the revisions. As the 10th revision was the most recent and updated revision of the ICD system, we decided to convert the codes of the 9th revision into revision 10. For this, using the corresponding possible conversion dictionaries and paying special attention to possible occurring mismatches, all the 9th revision codes in the MIMIC-IV dataset were converted. At the end of this first step, the MIMIC-IV dataset had only diagnosis codes of revision 10, and the eICU-CRD dataset was ready to be used, by using the already presented conversion of 10th revision codes.

Another crucial step in this process was to reduce the population of each of the datasets in order to obtain patients who were only associated with the most common comorbidities of migraine and migraine itself. When analyzing the population and their diseases in total, the work gets too overwhelming as there are many different diseases and associations between them. With more information comes more complexity that can be too hard to understand and navigate through. Patients with migraine are more often than not accompanied by a comorbid condition, adding to the complexity of the patient which can be explored. By narrowing down the population to only patients who expressed one of the conditions that can be seen in Table 3.4, the number of patients is reduced and it becomes possible to study in more detail how these comorbidities are related to each other. This facilitates the study of migraine, as it is a crucial step to identify the already known and most common comorbidities that are associated with it. The most common migraine comorbidities are known and studied at a level that allows to track them down, which allows the selection of patients with these conditions. The most common associations between migraine and other disorders include broader groups of pathologies, such as cerebrovascular dysfunctions, metabolic and endocrine comorbidities, epilepsy, sleep-related disorders, psychiatric disorders and pain syndromes, as well as gastrointestinal and immunological disorders. In Table 3.4, these groups of diseases are discriminated and their codification for the ICD system is presented for revision 10.

Taking into account each dataset in specific, the information of the total patients was filtered in order to only contain patients with comorbidities that are related to migraine. Filtering these patients allows for a reduction of complexity of patients, that would otherwise become too difficult to analyze if all the conditions were taken into account. With the prepared data of patients who present the comorbid conditions of migraine, it was possible to divide the total population into another group, of just patients who possess the migraine condition as a diagnosis. This means that the total population contains patients with comorbid conditions related to migraine, and the subgroup related to the migraine population, narrows down this population to only patients who exhibit the migraine condition and their comorbidities. At the end of this stage, it is possible to assess the data for patients with migraine or any of the comorbidities related to migraine.

Since there are some already studied gender differences among migraine patients, dividing the population based on their gender was also an important part of some of the performed analyses, as according to Peterlin et al. 2011. For each dataset, the total population can be divided into the two genders, and inside this total population, filtering for only the migraine patients, it could also be divided into the two groups by the corresponding gender. This allows us to understand which conditions take part in each of the gendered groups and reveal differences or similarities among each part.

At the end of all these stages and taking into account the two already prepared datasets of patients with migraine or with comorbidities related to migraine, or the group of migraine patients and their comorbidities, it was possible to assess the percentage of these conditions within both of these divisions.

Table 3.4: Comorbid conditions related to migraine and respective ICD codes of revision 10.

Broader descrimination	**Specific pathology**	**ICD codes**
Cerebrovascular dysfunction	Stroke	I63
Metabolic and endocrine comorbidities	Diabetes	E08-E13
	Obesity	E65-E68
	Insulin resistance	E88.81
	Hypothyroidism	E02, E03, E05
	Endometriosis	N80
Epilepsy	Benign occipital epilepsy	G40
	Benign rolandic epilepsy	G40
Psychiatric disorders	Major depressive disorder	F33
	Bipolar disorder	F31
	Post-traumatic stress disorder	F43.1
	Anxiety disorder	F40, F41
Other pain syndromes	Fibromyalgia	M79.7
	Chronic low-back pain	M54.5
	Pain accompanying dysmenorrhea	N94.6
	Temporomandibular disorder	M26.6
Sleep-related disorders	Insomnia	G47.0
	Sleep-disordered breathing	G47.30
	Restless legs syndrome	G25.81
	Narcolepsy	G47.4
	Advanced sleep phase	G47.22
	Parasomnia	G47.5
Gastrointestinal disorders	Periodontitis	K05.4
	Gastroesophageal reflux disease	K21
	Helicobacter pylori infection	B96.81
	Hepatobiliary disorders	K83.8
	Celiac disease	K90.0
	Irritable bowel syndrome	K58
	Inflammatory bowel disease (Crohn’s disease and ulcerative colitis)	K50, K51
	Constipation	K59.0
Immunological disorders	Multiple sclerosis	G35
	Systemic lupus erythematosus	M32
	Antiphospholipid syndrome	D68.61
	Primary Sjögren’s syndrome	M35.0
	Rheumatoid arthritis	M05
	Atopic diseases (eczema, asthma and rhinoconjunctivitis)	L20

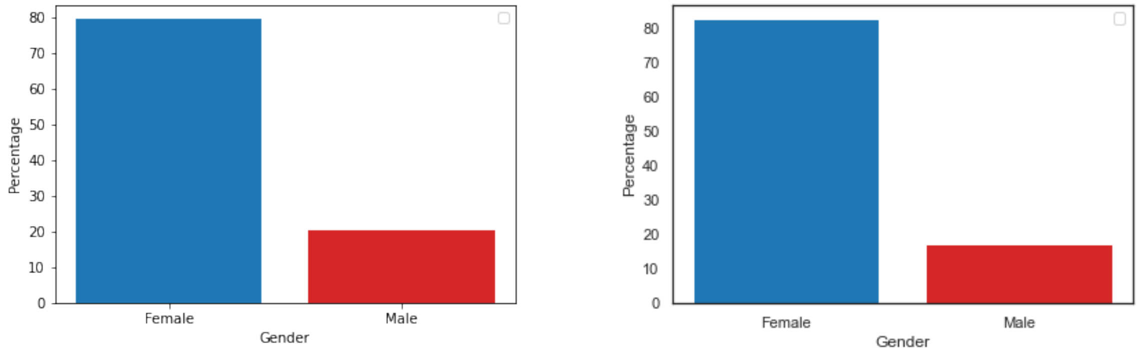
Chapter 4

Results and Discussion

In this Chapter, a characterization of the migraine patients within the available datasets was performed. The main results lay on the MIMIC-IV dataset, as it comprises more compact information about patients. Section 4.2 comprises the analysis of how the most common comorbidities of migraine are related to each other, using network representation and heatmaps with dendograms representations for the MIMIC-IV population. In Section 4.3, a clustering analysis is explored as to identify which are the subgroups and characterize the patients within said groups of migraine patients in the MIMIC-IV dataset. All of these analyses serve the purpose of deepening the knowledge related to these individuals. In Section 4.4 it is possible to see the simpler analysis done for the eICU-CRD dataset.

4.1 Characterization of Migraine patients

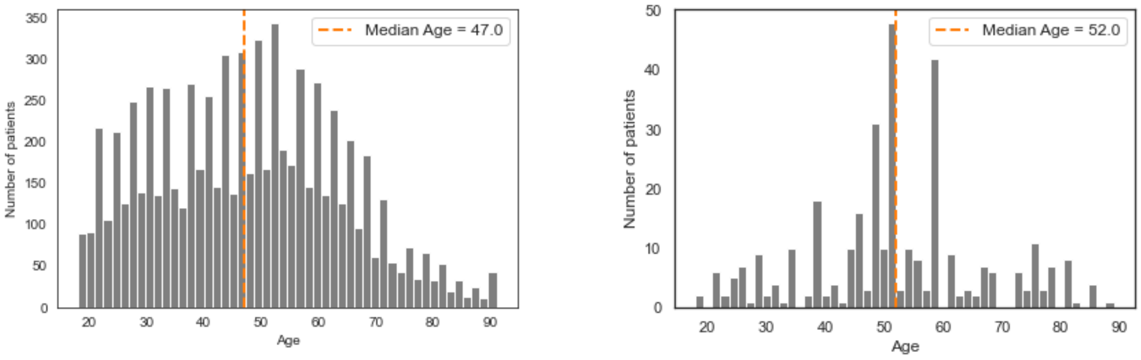
Migraine patients are complex subjects and account for a small percentage of the totality of patients in the studied datasets. With a number of 7 415 patients in the MIMIC-IV dataset, and only 117 in eICU-CRD, this group of patients display a set of characteristics that are intrinsic to the condition. Similarly to what can be found in the literature, women are more likely to suffer from migraine attacks than men up to three times. This follows the representation of the number of male and female patients who were associated with migraine in the MIMIC-IV dataset in Figure 4.1 (a). The greater percentage of patients who suffer from this condition can be seen in the 79.7% group of women, and 20.3% for men. As for the eICU-CRD dataset, both genders have a similar representation to MIMIC-IV in the migraine group of patients, a number of 57 men and 276 women. These patients have a median age of 47 years old for the MIMIC-IV dataset and an older median age for the eICU-CRD dataset of 65 years old, as it can be seen in Figure 4.2 (a) and (b) respectively. Another important feature that is innate to migraine patients is the distribution of how it affects subjects throughout their lifetime. The relationship between age and gender when it comes to migraine prevalence is shown in Figure 4.3. Showing consistency with the results from Vetvik and MacGregor 2017, when it comes to the MIMIC-IV dataset in (a), it is seen that the curves for the prevalence are similar to the ones in literature, which can be seen in Figure 2.1. Starting from the later teenage years, up until the 50-60 window, there is a noticeable higher percentage of female



(a) MIMIC-IV dataset

(b) eICU dataset

Figure 4.1: Percentages of male and female patients that are associated with migraine in the MIMIC-IV dataset in (a) and in the eICU-CRD dataset in (b).



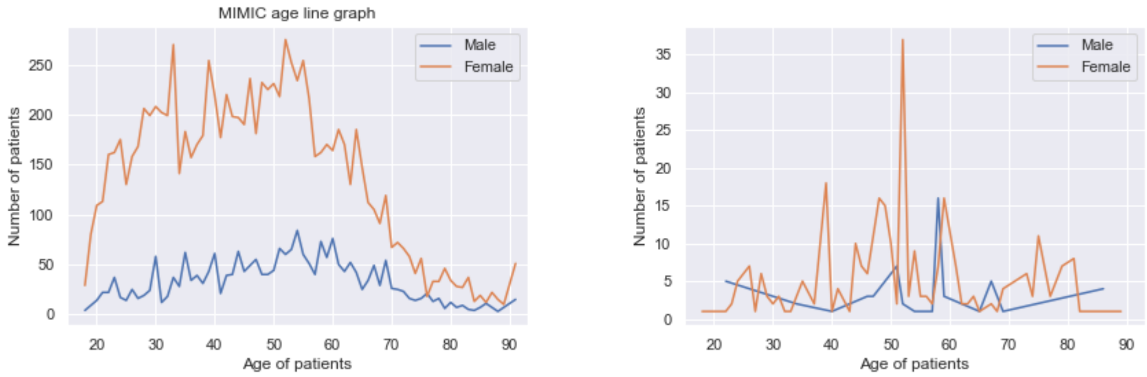
(a) MIMIC-IV dataset

(b) eICU dataset

Figure 4.2: Age histogram of migraine population in MIMIC-IV dataset in (a) and eICU-CRD in (b). It is possible to assess the median age in orange in each histogram.

patients suffering from migraine, falling into the fertile age period for women, confirming that the age at which the migraine attacks are more prevalent can be related to women's fertility window. It is most active when women enter their most fertile ages, from 20 years old to a noticeable decline right after menopause. As explained, migraine is thought to be commonly associated to hormones and the fertility period in which women menstruate in their lifetime, impacting the number of female patients that suffer from this condition. Thus, it makes sense that there is an increase in the total amount of female patients who suffer from migraine at these stages of life. As for men's age distribution, it seems steady with no major differences, and declining as predicted in the literature, in the later stages of life. Regarding the eICU-CRD dataset in Figure 4.3 (b), the network portraits a different scenario from what is shown in (a). However, this can be explained through the fact that eICU-CRD contains information about critical care patients. In critical care units, patients are in a completely different context than what is analyzed as the whole population. Meaning that male and female patients in the migraine subgroup of this dataset have a heterogeneity in the distribution of age, understandable through the context in which patients are inserted.

Since this dissertation has a focus on the most common comorbidities of migraine patients, it is im-



(a) MIMIC-IV dataset

(b) eICU dataset

Figure 4.3: Age distribution graphs representing the number of patients with migraine-related diagnosis divided by gender and ages for the MIMIC-IV in (a) and eICU in (b) datasets.

portant to understand the prevalence of said conditions inside this group. This can be seen in Figure 4.4, where the percentage of each of the most common comorbidities can be evaluated. The most prevalent disorder among migraine patients is gastroesophageal reflux disease, with a percentage of 35.91%, followed by anxiety disorder with 33.93%. Obesity (20.02%), diabetes (16.02%), hypothyroidism (15.31%), constipation (13.74%), and insomnia (10.47%) have a prevalence above the 10%, meaning that these are also significant disorders within this set of patients. On the other hand, there were some conditions that did not partake in this group of subjects, having a 0% of prevalence in the population, namely chronic low back pain, periodontitis, parasomnia, advanced sleep phase, narcolepsy, and sleep-disordered breathing. Some other conditions that had a lower appearance were rheumatoid arthritis (0.07%), atopic diseases (0.09%), and insulin resistance (0.25%). Having a broader understanding of how each of these diseases affects the migraine population in the MIMIC-IV dataset, evaluating how these disorders are connected to each other can be of benefit and that step is performed in the next section. Taking into account the presented results of less than 0.01% for conditions such as chronic low back pain, periodontitis, parasomnia, advanced sleep phase, narcolepsy and sleep-disordered breathing, these disorders were removed from the dataset and they will not be further analyzed.

4.2 Correlation between Migraine comorbid conditions in MIMIC-IV patients

Being migraine a complex condition that can be associated with a wide spectrum of other disorders, as previously seen in Chapter 3, an analysis of how the most commonly occurring comorbidities within this group of patients co-occur was performed. Taking into account the conditions in Table 3.4 and relating them to the prevalence that was found among patients, it was possible to analyze which ones are most commonly associated with each other. This allows studying how important these relations between disorders inside the migraine population are and how they differ from the total MIMIC-IV population of patients with these comorbidities. Thus, this allows us to understand if the migraine condition partakes

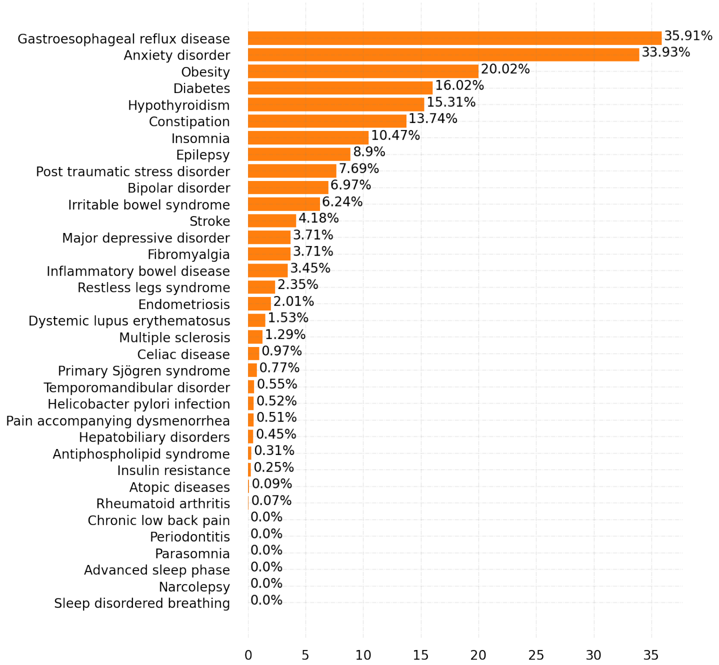


Figure 4.4: Comorbidities related to migraine and their percentage in the subgroup of patients that are affected by this condition.

in the degree to which these comorbidities co-occur in the population or, on the other hand, if there are no differences found. In order to do this, the initial step is to perform an analysis of the correlation between the comorbidities of migraine and how their occurrence was to quantify how these comorbidities take place within each individual patient: combining the disorders in pairs and counting how many times the combination occurred among patients, allowing us to understand the relevance of these associations. Two criteria were chosen in order to facilitate the process and divide the population into smaller groups and get a more in-depth study. Computing the sum of how many times each combination of two comorbidities appeared within all patients, it was also seen how many of these combinations occurred within the migraine patients group and dividing it beyond that, the conditions were separated into gender specific groups. Thus, a final count of how many times these pairs of the comorbidities appeared in the total population, total female population, total male population, and inside each of these gender groups: total female migraine population and total male migraine population. This allowed to understand which pair of conditions were more connected to each other and which pair happens more commonly inside these restricted groups of patients.

4.2.1 Network visualization of comorbid conditions co-occurrences

In order to visualize these connections between disorders, networks were built resorting to Python's *NetworkX*[1] package and the software Gephi[2] for visualization purposes, which can be seen in Figure 4.5. The nodes represent the conditions found in the Table 3.4 and the interactions between each other, meaning how many times they co-occur together, can be seen through the edges. The wider the edge,

[1] https://networkx.org/

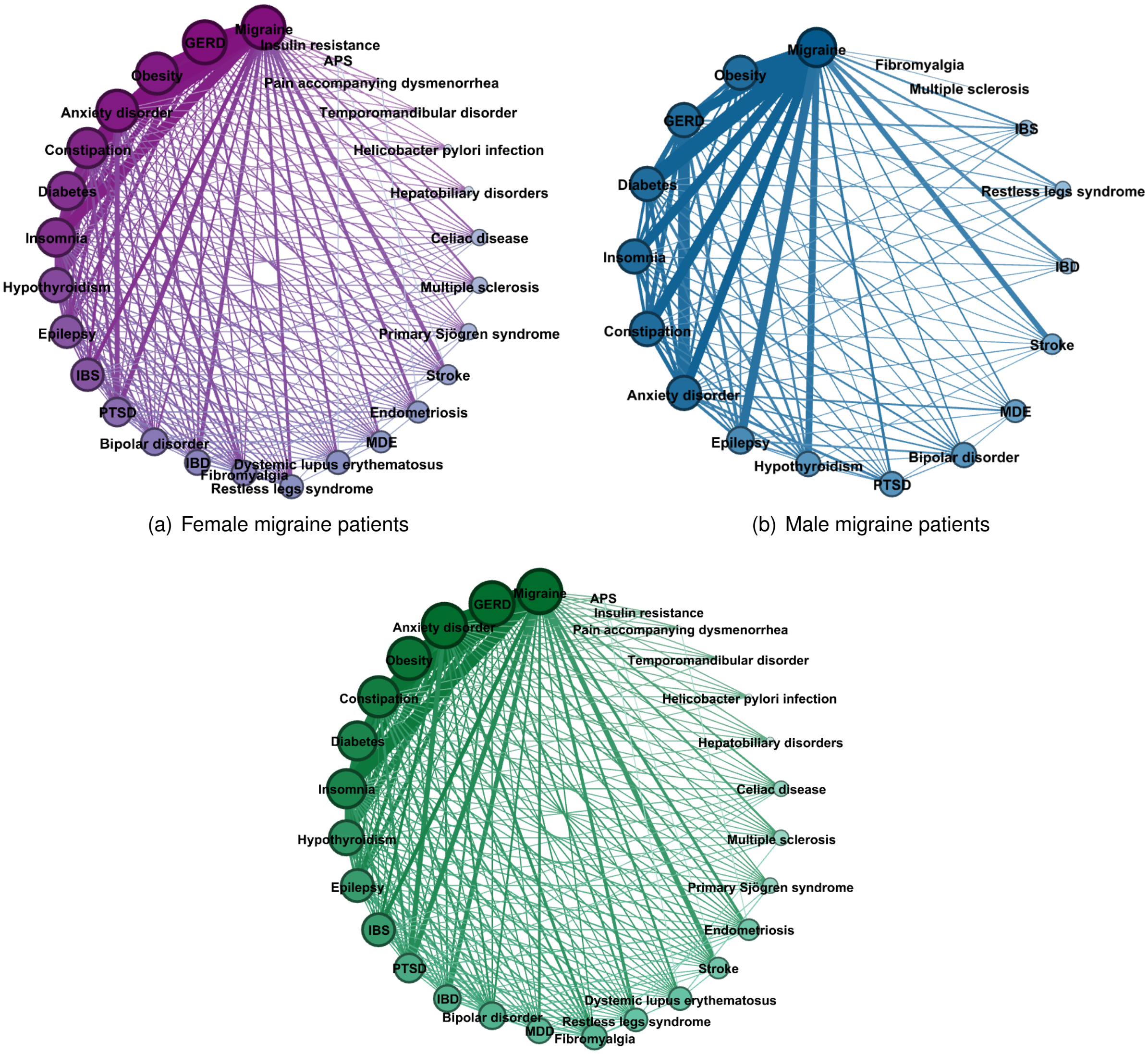
[2] https://gephi.org/

the more times the combination of these two conditions occurs, highlighting how strongly they are connected to each other. One important adjustment done at the time of obtaining the graphs was to only accept nodes whose weight accounts for at least 1% of the population. Hereby, the combinations of the two comorbidities had to account for at least 1% of the combinations in these graphs. The graphs that are presented have been adjusted to display a circular layout, and Gephi's software supports computing the degree of these nodes, making it possible to order them through this measurement and in a counterclockwise direction. The presented nodes were adjusted so that the node size, as well as the edge size, are directly correlated with this measure. The edges account for how many times the combination of the disorders occurs, taking into account the edges' weight.

For the MIMIC-IV population, it was possible to verify through Figure 4.5 (c), the totality of the migraine population, that the conditions which are more highly co-related to migraine are gastroesophagal reflux disease, anxiety disorder, obesity, constipation, diabetes, and insomnia. These conditions present a strong edge between each other and are the last ones in the circular graph, as opposed to other conditions presented at the beginning of the circular graph, such as antiphospholipid syndrome, insulin resistance, and pain accompanying dysmenorrhea. This is directly related to the fact that these disorders appear to have a lower presence in patients with migraine, as previously seen in Figure 4.4, where these conditions presented a percentage lower than 1%. Since their appearance in this population is low, it is clear why they do not partake a relevant role in this graph.

Regarding the differences between the female and male populations, it is demonstrated through Figure 4.5 (a) and (b) that there are some visible changes when dividing the population by gender. The most apparent difference is the number of nodes contained by each graph. For women, in Figure 4.5 (a), the network is similar to the whole population of migraine patients, with the same number of 28 nodes. However, for the network related to the male migraine population, in Figure 4.5 (b), the number of nodes is reduced to 18. This reduction of nodes in men's migraine population can be explained through the fact that some of the conditions that are seen in women are not present in this subgroup, since these are related to the female reproductive organ, and the menstrual cycle. Endometriosis and pain accompanying dysmenorrea can be seen in women's graph, but not in men's. This induces a clear difference when it comes to the results of these graphs. Apart from that, what is seen to be distinguishable between both genders can be related to the order in which the nodes are presented. In men, migraine is highly correlated to diabetes, obesity and gastroesophageal reflux disease. While in women, although the order at which these nodes are presented changes, changing the importance of said connections, it is seen that conditions such as anxiety disorder and hypothyroidism account for those differences. Overall, the most prominent conditions that are highly associated with migraine are gastroesophageal reflux disease, obesity, diabetes, anxiety disorder, insomnia and constipation, confirming what could be seen through the migraine population prevalence network in Figure 4.4.

For the purpose of analyzing more in-depth the computed networks, the most common measurements and their values for each of the graphs can be seen in Table 4.1. One of the most important measures and the one that was used to order and size the nodes through the circular display is the degree. In this case, we can see that each of the nodes in the female patients' network in (a) and the



(a) Female migraine patients

(b) Male migraine patients

(c) Total migraine patients

Figure 4.5: Circular graphs representing how the most common comorbidities of migraine are related to each other in MIMIC-IV's migraine patients. Nodes represent the disorders, and the edges represent the times each combination of two occurs. (a) contains 28 nodes, 231 edges; (b) contains 18 nodes, 97 edges; (c) contains 28 nodes, 234 edges. GERD - Gastroesophageal reflux disease; IBD - Inflammatory bowel disease; IBS - Irritable bowel syndrome; MDE - Major depressive episode; APS - Antiphospholipid syndrome; PTSD - Post traumatic stress disorder.

total patients in (c) have a similar average degree, meaning that for these two graphs, the nodes are connected to a similar number of edges. When it comes to the average path length, the highest value is associated with the women's network, and the lowest value is for the male's network.

Table 4.1: Measurements for the obtained graphs related to total migraine patients, female and male migraine population.

Features	**Female patients graph**	**Male patients graph**	**Total patients graph**
Average degree	16.5	10.78	16.714
Network Diameter	2	2	2
Average path length	1.389	1.366	1.381

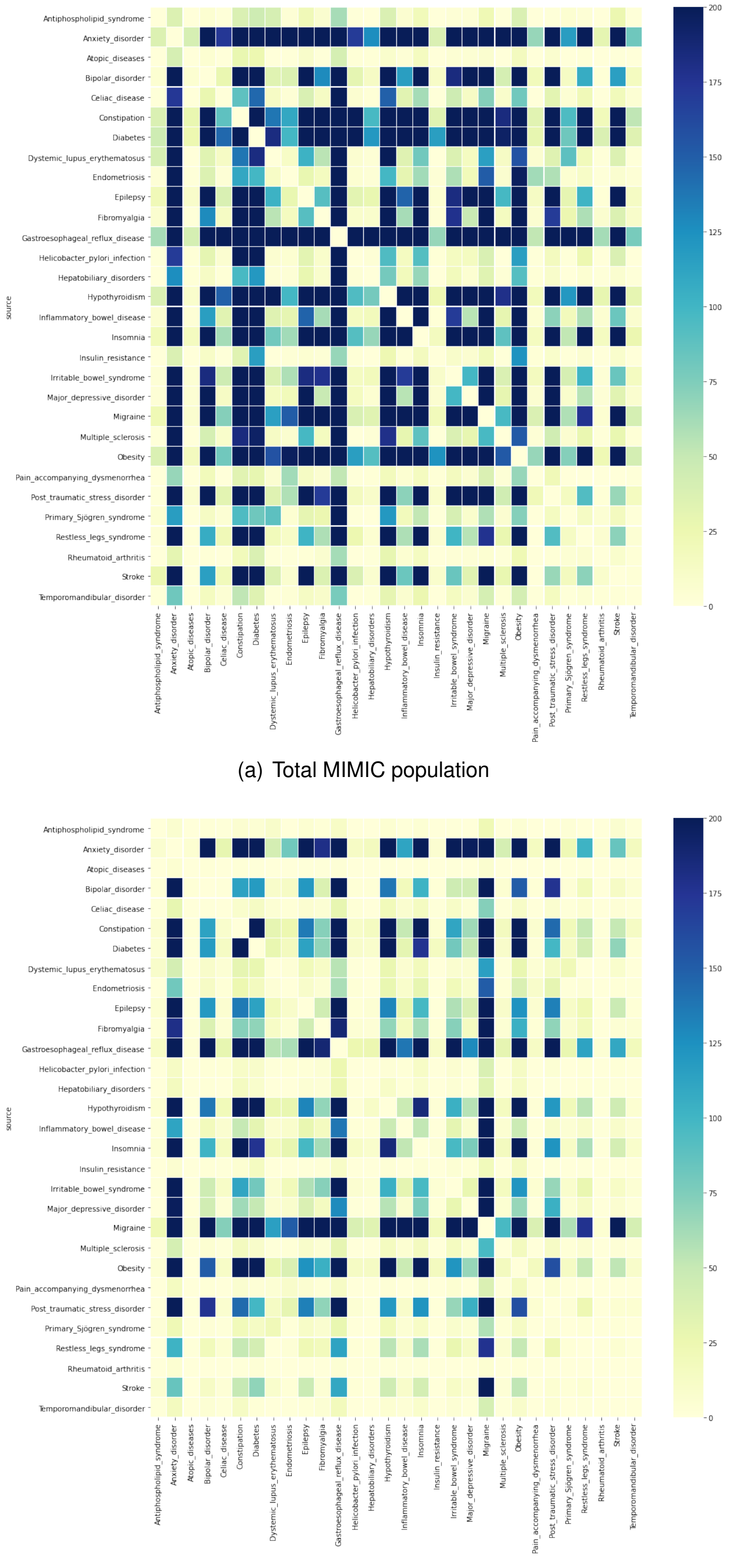
4.2.2 Heatmap visualization of comorbid conditions co-occurrences

To visualize more in-depth the obtained results in the previous section and further understand how these comorbid conditions are associated with each other, heatmaps were computed while taking into account how many times each condition appears together with another condition within the patients. Firstly, and through Figure 4.6 (a), it is possible to assess how often these conditions appear together in the whole population of MIMIC-IV. Within this total population, a condition that stands out and can be seen to have a correlation between a high number of disorders is gastroesophaseal reflux disorder, as well as anxiety disorder, constipation, and diabetes. Hereby, patients in the total dataset of MIMIC-IV have high association with these comorbid conditions. Filtering this population into the desired set of individuals that present migraine, it can be verified through Figure 4.6 (b) that migraine has the highest number of counts among all the pairs of the comorbid conditions. This makes sense, as all patients display this disorder and confirms that these comorbidities are related to each other.

Dividing the MIMIC-IV population into female and male patients, allows us to understand which conditions play a more significant role in each of these patients. As previously done through the networks, in the computed heatmaps that account for how many times two disorders are seen together on a patient, it is possible to see differences between how they appear in the total female population in MIMIC-IV, and differently from the total female migraine population. This can be analyzed through Figure 4.7 (a) and (b) respectively. These two heatmaps showed a similar outcome to the total population ones, seen in Figure 4.6. As for the male population, it can be verified through Figure 4.20, that although there is a reduction in the number of associated comorbidities, the heatmap shows that some of the already seen conditions such as anxiety disorder, constipation and diabetes are important in the total population. For the subgroup of the male migraine population in (b), there is a noticeable change from the counterpart of the female migraine population. All the associations between the comorbidities have a smaller count among men, and the values are lower overall.

4.2.3 Statistical relevance of comorbid associations

The analysis of the counts of how these pairs of comorbidities occur within the migraine population, it can be difficult to understand their real significance and importance. Taking into account the migraine population, this means that seeing a high number of times two conditions co-occur may not mean that these two conditions are significant within the migraine population, because they may also have a high number of occurrences for the totality of the MIMIC-IV population. On the other hand, two conditions that occur but at a lower rate when compared to other more common co-occurrences, may be overlooked and deemed as not as important, when in reality their co-occurrence is of significance in the migraine population when compared to the totality of the MIMIC-IV individuals. Thus, it was necessary to understand what is the statistical relevance of said occurrences. For this, it was possible to compute the p-values associated with each of the occurrences of conditions coupled into pairs. For each patient, it was seen which conditions it was associated with, and within each patient, combining each of the conditions into pairs and counting the number of times this combination of conditions appeared within the

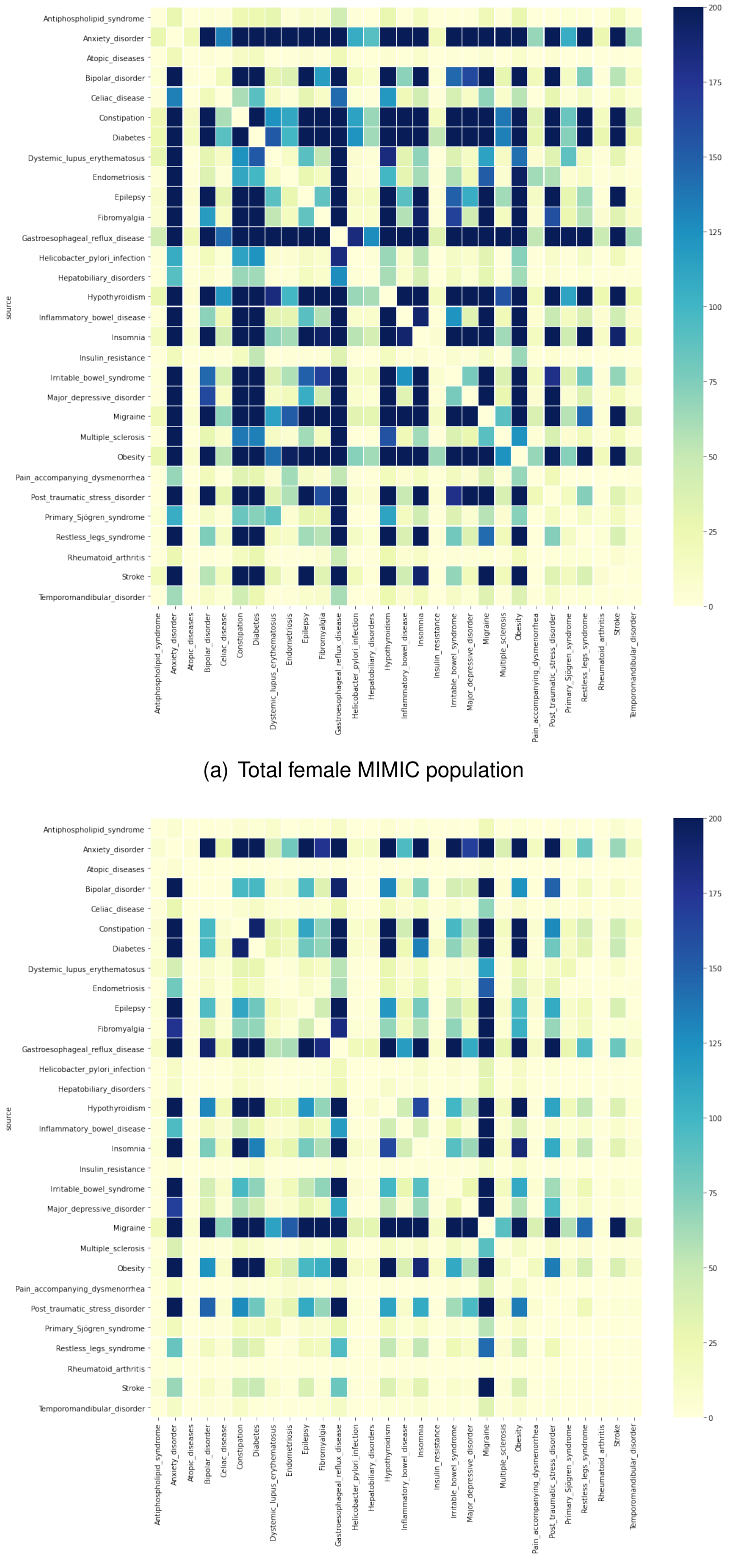


(a) Total MIMIC population

(b) Total migraine population

Figure 4.6: Heatmap displaying the counts of two of the comorbid conditions associated to migraine, in this case for the whole MIMIC-IV population and the migraine subgroup. Darker shade of blue codes for a higher appearance, opposed to lighter shades of blue/white, coding for lowest values.

population. Taking into account all the possible group divisions for the MIMIC-IV population, that is the totality of patients, migraine patients, and dividing these two groups by gender, the counts were found

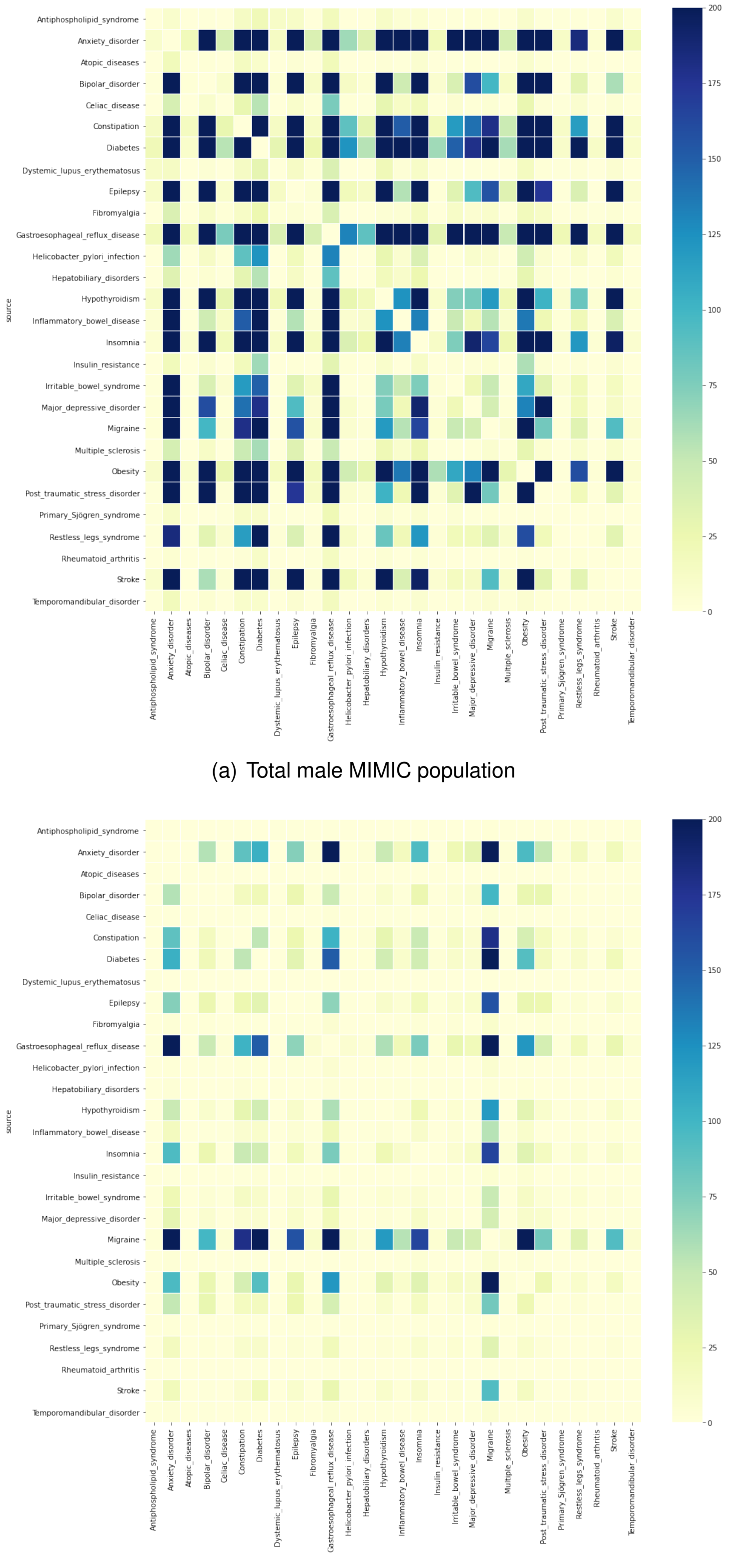


(a) Total female MIMIC population

(b) Total female migraine population

Figure 4.7: Heatmap displaying the counts of two of the comorbid conditions associated with migraine, in this case for the whole **female** MIMIC-IV population and the **female** migraine subgroup. Darker shade of blue code for a higher appearance, as opposed to lighter shades of blue/white, coding for lower values.

for each of the groups. The p-value of each combination of the two comorbid conditions was obtained



(a) Total male MIMIC population

(b) Total male migraine population

Figure 4.8: Heatmap displaying the counts of two of the comorbid conditions associated with migraine, in this case for the whole **male** population and the **male** migraine subgroup. Darker shade of blue codes for a higher appearance, as opposed to lighter shades of blue/white, coding for lower values.

by resorting to the hypergeometric function of the *SciPy* Python's package. [3]This function enables the

[3] https://scipy.org/

calculation of the p-value, looking into the probability of a pair of conditions occurring within the different divisions that were just mentioned. This analysis aids in the understanding of the possible differences in the population regarding which conditions are more or less often associated with each other, giving an insight that can be helpful to decide what next steps to take. It was pertinent to build networks while narrowing to pairs of associations that are associated with a p-value that is lower than 0.05. This allows for the understanding of which conditions are more associated with each other in each of these groups and subgroups that have been formed.

4.2.4 Network visualization of relevant comorbid associations

Similar to how the networks were obtained in Section 4.2.1, the graphs for the obtained combinations of two comorbid conditions were obtained, taking into account the p-value of at least 0.05 for each of them. The layout that was chosen for this representation was Force Atlas 2 and nodes were divided resorting to the modularity feature of Gephi, taking into account the weights. The nodes represent the conditions and the degree represents how many times each combination occurred. It was also necessary to filter out combinations of conditions that did not account for at least 1% of the population.

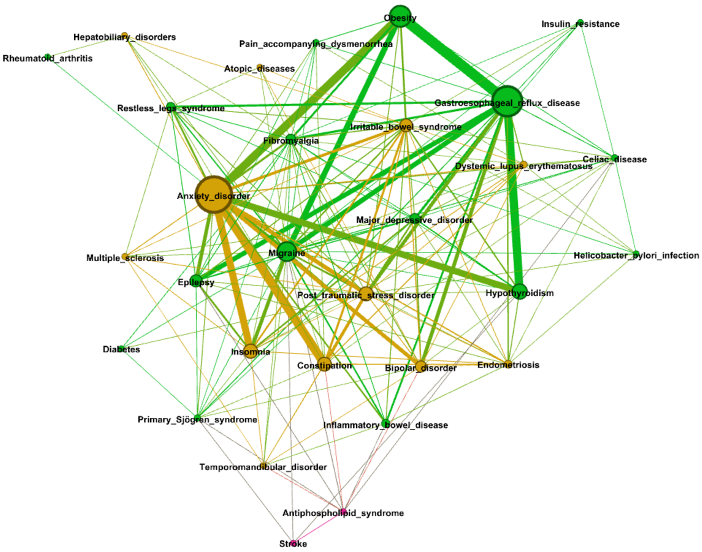
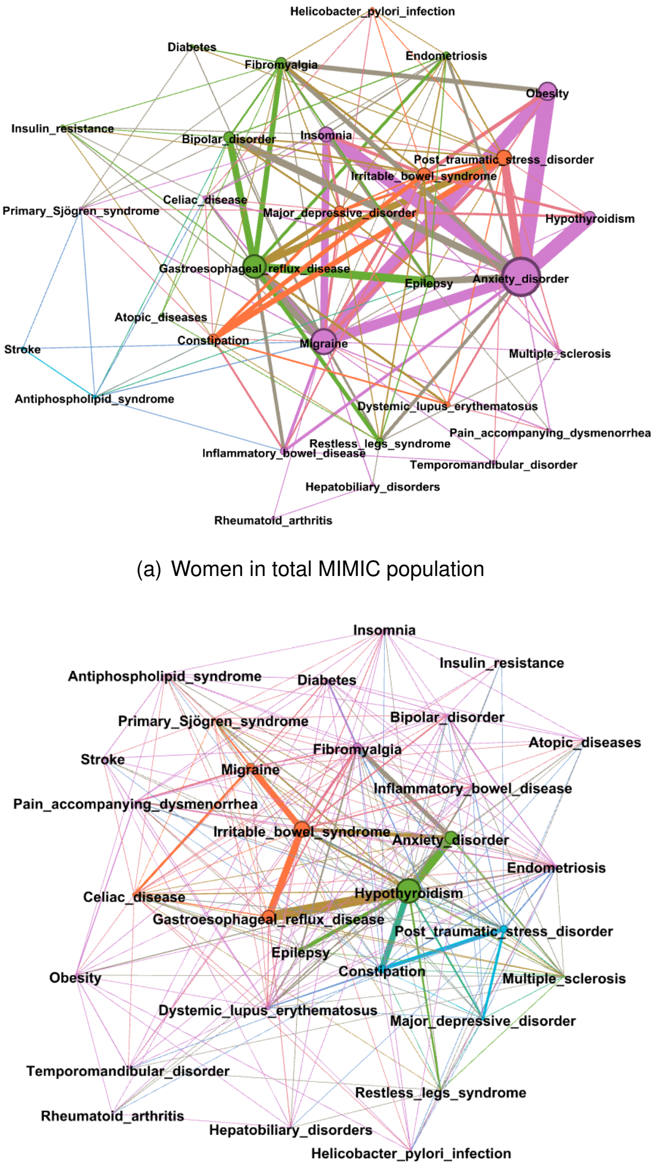


Figure 4.9: Representation of the graphs containing the most relevant combinations of comorbididities within the **total** population of the MIMIC-IV dataset and the **total** migraine population, which stated a p-value > 0.05. Contains 30 nodes, 177 edges.

Regarding the network computed for the total population of migraine when comparing to the MIMIC-IV population who are associated with the comorbid conditions of migraine, we can verify in Figure 4.9, that for the most part, there are three groups in which the comorbidities are related to each other. The biggest node is represented by an anxiety disorder in yellow, which connects heavily to conditions in this group such as insomnia, constipation, bipolar disorder, post-traumatic stress disorder, irritable bowel syndrome, and endometriosis. The nodes that are colored green are associated to the biggest node which is gastroesophageal reflux disorder, connected to hypothyroidism and in which migraine is a part of, as well as epilepsy, major depressive episode, and obesity. There is another group of nodes with the color pink that are nodes with a size smaller when compared to the ones in yellow and green that were just mentioned. Further analyzing this, for migraine patients there are two conditions that stand out and are significant when compared to what is represented in the total MIMIC-IV population, which are anxiety disorder and gastroesophageal reflux disease.

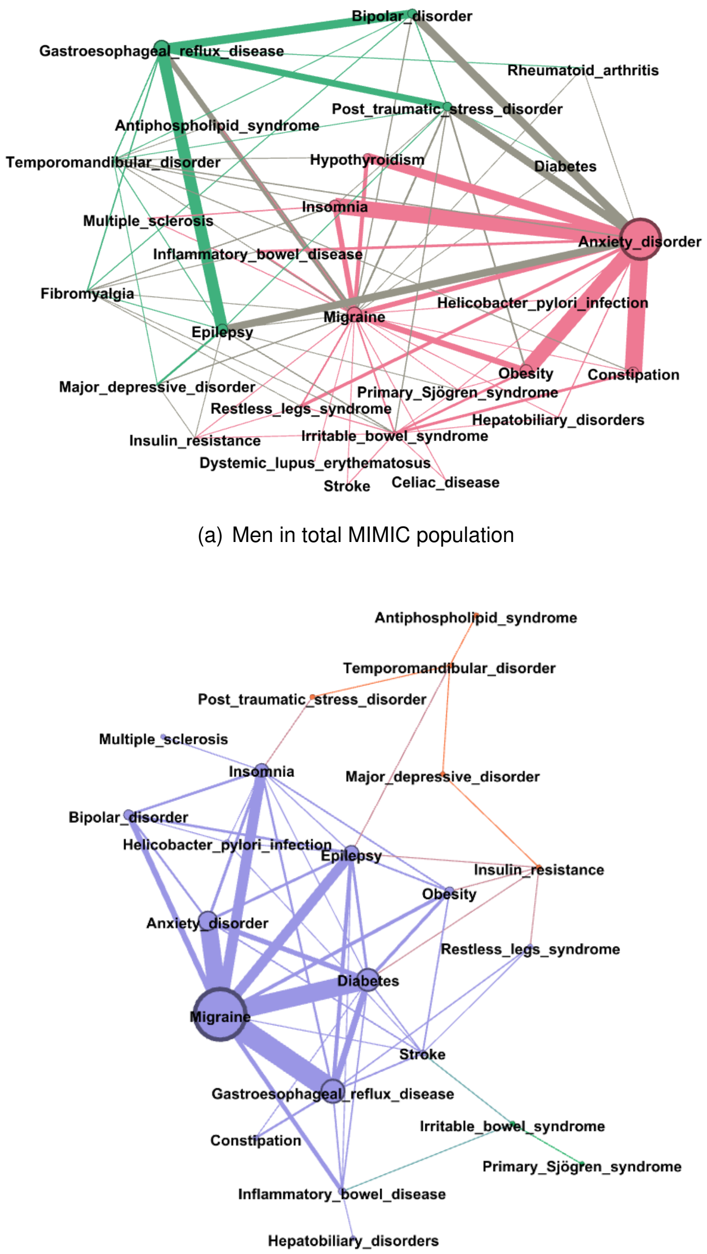


(a) Women in total MIMIC population

(b) Women in migraine subgroup

Figure 4.10: Representation of the graphs containing the most relevant combinations of comorbididities within the **female** population MIMIC-IV dataset and the migraine **female** population, which stated a p-value > 0.05. (a) Contains 30 nodes, 143 edges. (b) Contains 30 nodes, 229 edges.

Regarding the networks that contain the female population, it is possible to verify a big difference in relation to the whole female population of MIMIC-IV and the migraine female subgroup. In Figure 4.10 (a), it is possible to see that some of the common comorbidities of migraine are often coupled together, even when the totality of the population is not always associated to migraine. Female patients in MIMIC-IV often present the anxiety disorder condition together with obesity, insomnia and at last with migraine. This means that these anxiety disorder and anxiety are seen together in the totality of the female patients



(a) Men in total MIMIC population

(b) Men in migraine subgroup

Figure 4.11: Representation of the graphs containing the most relevant combinations of comorbididities within the **male** population MIMIC-IV dataset and the migraine **male** population, which stated a p-value > 0.05.(a) Contains 27 nodes, 81 edges. (b) Contains 22 nodes, 54 edges.

in this dataset. It is also possible to verify that because the presented conditions are the most common comorbidities of migraine, that they are, as expected, highly related to each other. In the total MIMIC-IV population, it is possible to verify that migraine is often paired with a variety of conditions, which can be seen through the connections in pink. Comparing it with the subgroup of migraine, it is possible to identify much broader connections in subgroups of Figure 4.10 (b). This means that all the conditions that can be seen in this network are very much in association with each other within the migraine female group, meaning that these patients are complex ones, with many associations of conditions related to migraine comorbidities. The most significant connections lay between migraine, irritable bowel syndrome, anxiety disorder, constipation, gastroesophageal reflux disease and post-traumatic stress disorder. Women are shown to have a great number of significant co-occurrences among the already known most common comorbidities.

As for the men in the MIMIC population, it is possible to observe in 4.11 that the way comorbidities interact with each other differs from the group they are inserted in. In men's migraine subgroup, we can see that there are fewer associations that can be relevant when compared to the whole male population of men in MIMIC-IV. Meaning that although migraine plays a part, most common comorbidities are also often together even when migraine is not a diagnosis for patients and thus not as relevant. In 4.11 (b), we can see that the most common associations of disorders between male migraine patients are between the migraine condition and disorders such as gastroesophageal reflux disease, diabetes, anxiety disorder, insomnia and epilepsy. Through the two different colors present in this, it is possible to see that these are more often co-occurring with the migraine condition and associated strongly with each other. On the other end, the group which is colored in pink can be seen to have fewer connections to the migraine condition itself but to the comorbidities that are around this condition, although the nodes and edges are not heavily marked. As for the totality of male patients in the MIMIC-IV dataset, we can see that in (a) that anxiety disorder has a high prevalence in the population, being the biggest node in size and the one that connects to a high number of other nodes heavily. There is a visible group of these diseases who are highly related to each other and can be found in the color pink. This means that the male population of MIMIC-IV who has at least one of the comorbidities associated with migraine, has a high percentage of male patients who show anxiety disorder, obesity, constipation, insomnia and hypothyroidism, as previously seen in the previous Sections. The group in green is related to less frequent combinations but gastroesophageal reflux disease is seen to be highly connected to epilepsy among male patients in MIMIC-IV.

4.2.5 Dendogram visualization of comorbid associations

In order to understand these connections even better and see how they occur in the population, it was possible to obtain dendograms associated to heatmaps. There are several metrics that can be used to obtain the dendograms with heatmaps. The results depend, of course, on our desired outcome, meaning that what we looked for was the closeness of associations to the principal condition that is migraine. For this situation, the metric that was used for all the dendograms and heatmaps was the euclidean, with the Ward's method. This type of metric allowed for the best outcome compared to all the other metrics, meaning that the condition migraine showed the best results in terms of coupling together to other conditions with this type of metric.

As seen in Figure 4.12, through the association of heatmaps to the dendograms, it is possible to visualize which are the most relevant conditions and which ones deem to be in closer relationship with migraine, regarding the population as a whole. The totality of the associations of the comorbidities related to migraine in MIMIC-IV patients total population shows that migraine is highly related to two other conditions, as seen in the dendogram: irritable bowel syndrome and endometriosis. This latter condition is related to women's reproductive organ, and accounts for a significant association between the other conditions, as well as migraine. This association between migraine and endometriosis follows what has been hypothesized, because the migraine condition has been proven to be associated with menstruation and the hormone cycle of women in their fertile era. Apart from the women-specific related condition and irritable bowel syndrome, this group of three disorders is connected to a larger group that contains conditions such as fibromyalgia, post-traumatic stress disorder, pain accompanying dysmenorrhea, celiac disease, rheumatoid arthritis, antiphospholipid syndrome, insulin resistance, dystemic lupus and temporomandibular disorder. These conditions have a greater likelihood of being associated to each other, as the group seen in the dendogram, being the most likely conditions to be associated to migraine and to each other in pairs. It is possible to also see that all conditions are highly associated to the migraine condition, but not all of them are related to each other. Namely, the ones that appear on the top left of the heatmap, with p-values that are high and thus not as relevant.

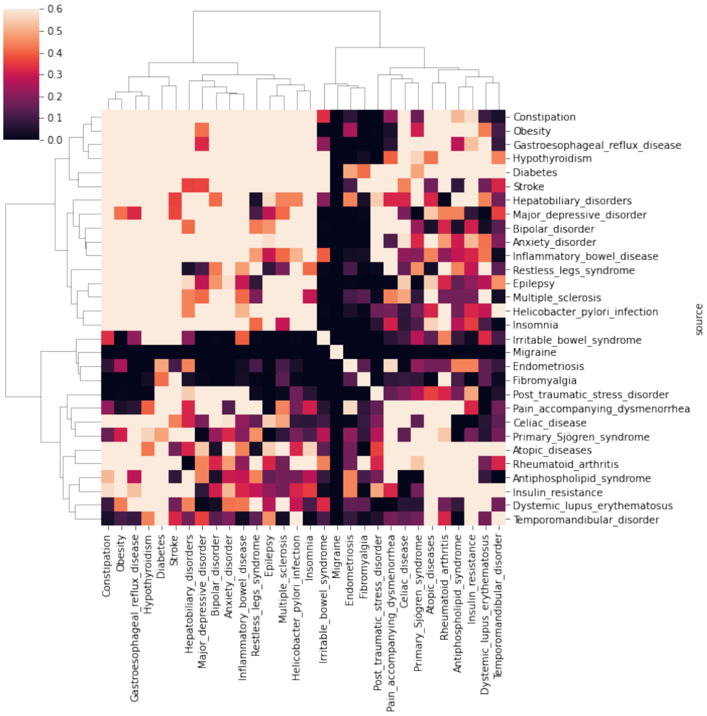
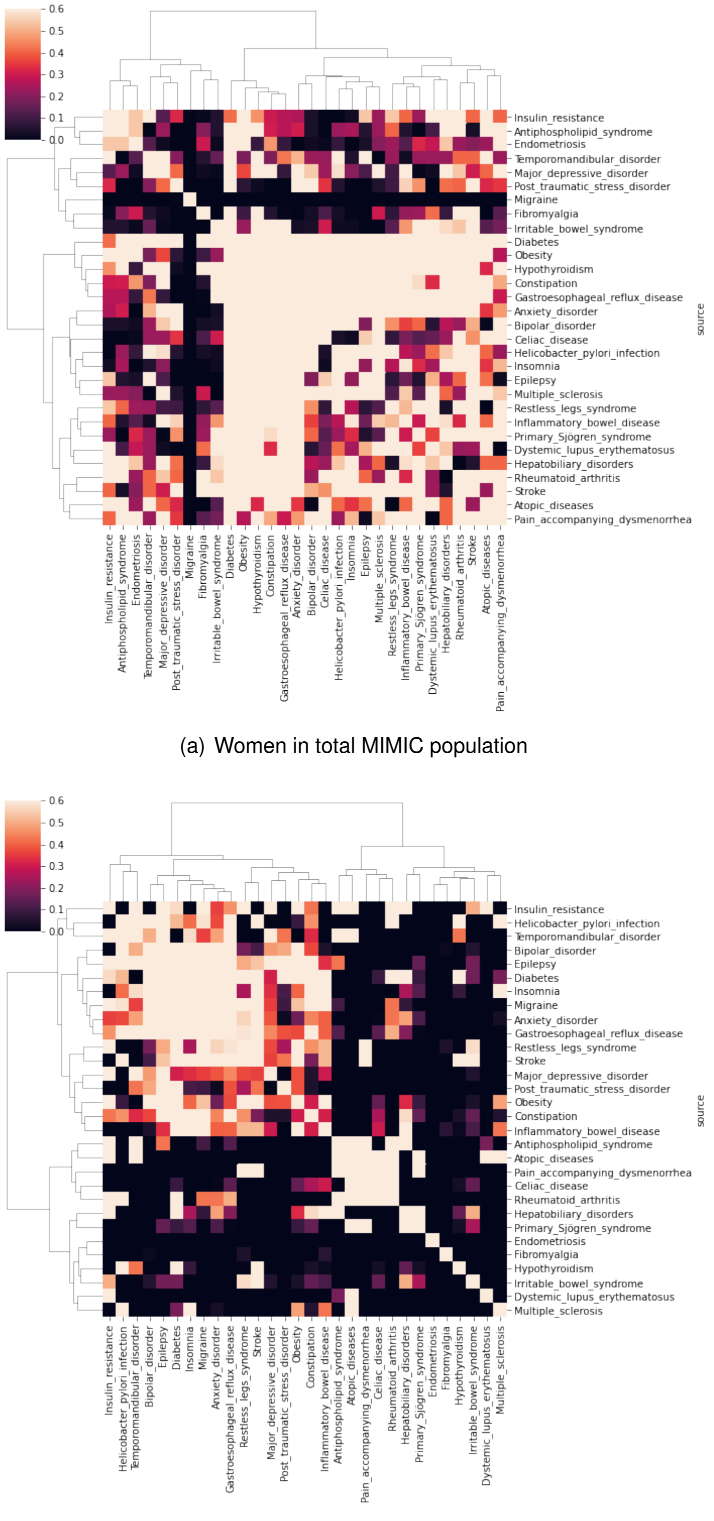


Figure 4.12: Dendogram with a heatmap of the **total** migraine population and displaying the probability of a patient having two of the comorbid conditions associated to migraine. A darker color codes for a lower p-value, thus a higher statistical relevance of the association between two conditions.

When it comes to the gender basis analysis, it was possible to have two different outcomes regarding each gender. One of the analysis is done based in the whole MIMIC population that is associated to any of the comorbid conditions. Another analysis is based on the population of these said patients with comorbid conditions related to migraine, but that are also associated to migraine, meaning that we have the migraine population and their associated comorbidities. When analysing the dendograms with heatmaps related to women and their comorbid associations, it is possible to see in Figure 4.13 (a) that the most relevant conditions follow a similar trend as the population as a whole, in which the migraine is associated to irritable bowel syndrome and in addition to that, it also is linked to fibromyalgia. This means that for the totality of female patients in MIMIC-IV, there are associations of the comorbidities that are common and thus not as relevant to the context in which they are inserted into. However, it is possible to verify that migraine is highly related to all the comorbidities found, with low values of p-values and a darkest columns in the heatmap.

As for Figure 4.13 (b), this is related to women in the migraine subgroup, hinting to all the comorbid conditions that are associated and co-occur in migraine patients when focusing on women. There is a clear difference between the heatmaps presented for the whole female population and when filtering only for female migraine population. There is a greater number of conditions that are highly associated with each other in a significant manner, thus the heatmap is colored in darker colors. Inside the female population that has migraine condition, it is possible to assess that conditions such as hepatobiliary disorders, primary Sjogren syndrome, endometriosis, fibromyalgia, hypothyroidism, irritable bowel syndrome, dystemic lupus erythematosus and mutliplesclerosis are highly connected to each other. Meaning that these connections are significant for the female population who presents migraine. This group of disorders is also connected to another group which is comprised of antiphospholipid syndrome, atopic diseases, pain accompanying dysmenorreha, celiac disease and rheumatoid arthritis. As it can be verified, since this group of migraine patients comprises female patients only, it was expected and verified that conditions associations to women's reproductive organ would be significant, such as endometrisosis and pain accompanying dysmenorrhea. It is also important to point out that some of the psychological disorders that were seen in the total population, have not been considered as relevant inside the migraine population and considering only women. Since these patients are necessarily associated to migraine, it is possible to observe that this condition shows less relevant associations to other conditions, and carries less value than other co-occurrences in the female population of migraine patients.

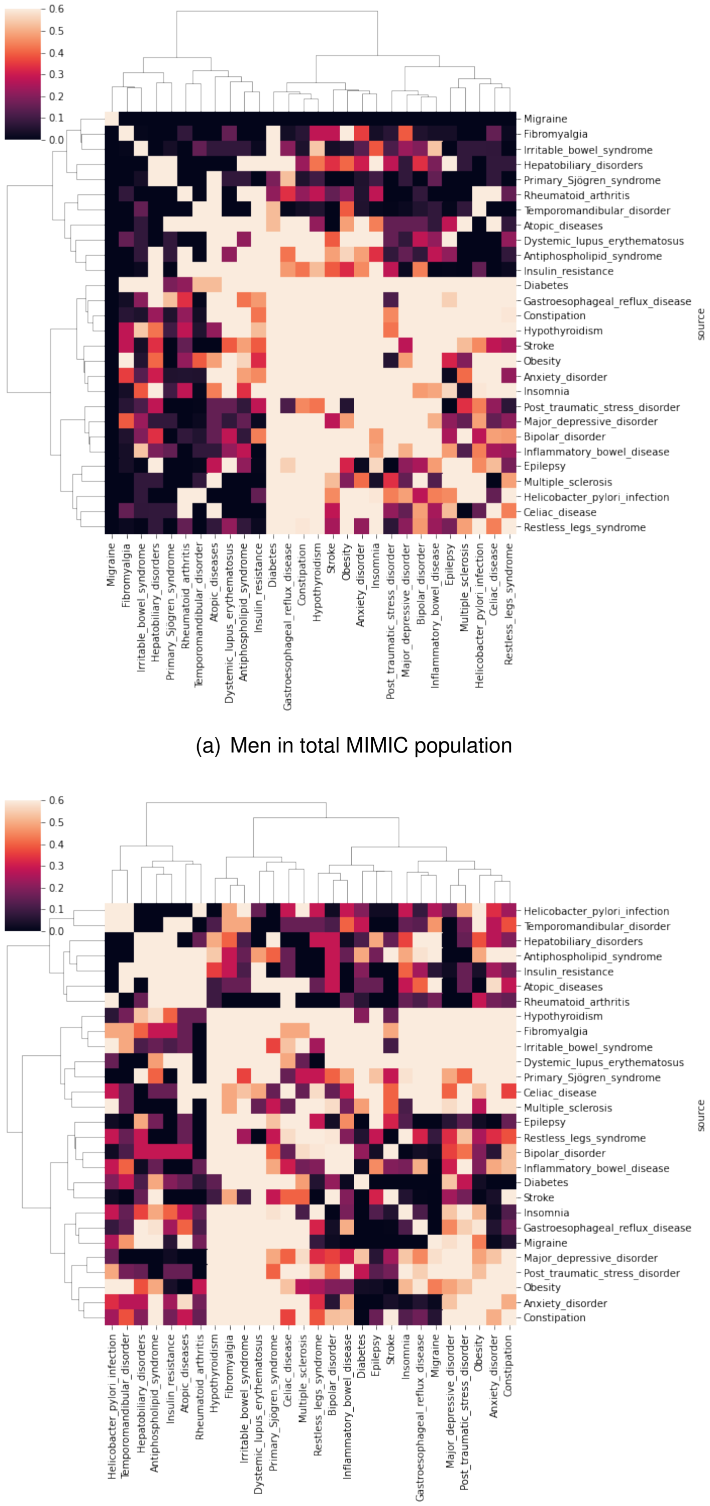
As for the male population, as seen in Figure 4.14 (a) and (b), the most prominent correlations to migraine are different from what is seen in the female population's dendograms with heatmaps. For the totality of male patients within MIMIC-IV, who present at least one of the most common comorbidities of migraine, it was possible to verify that in this group, migraine is associated to conditions such as Fibromyalgia, Irritable bowel syndrome, Hepatobiliary disorders, and Primary Sjögren's syndrome. It is also important to verify that migraine is significant to all the associations between all the conditions, as they are frequent comorbidities which makes sense. However the relations in women can be seen as a group less relevant when looking into the female population without discriminating the migraine condition. Men seem to have a wider range of conditions associated to it but not necessarily when possessing the migraine condition. In this case, there are no associations to the women's reproductive system diseases as they do not apply in this case and would not be relevant to show. Looking into Figure 4.14 (b), it is possible to see a greater difference when comparing to the migraine female subgroup of patients. The p-values are higher in most conditions, meaning that the occurrence of the pairs of conditions are not as relevant and specific to the migraine subgroup. Migraine is grouped together with psychiatric disorders such as major depressive disorder, post-traumatic stress disorder, anxiety disorder, meaning that male patients who present the migraine condition may also be experiencing these conditions. Less relevant associations between conditions can be traced to hypothyroidism, Fibromyalgia, Irritable bowel



(a) Women in total MIMIC population

(b) Women in migraine subgroup

Figure 4.13: Dendogram with a heatmap displaying the probability of a patient having two of the comorbid conditions associated to migraine, in this case for women in MIMIC-IV. A darker color codes for a lower p-value, thus a higher statistical relevance of the association between two conditions. In (a) we have the p-values associated to **women** in relation to the whole MIMIC population and their comorbid relations. In (b) we have the p-values for women's comorbidities correlations between **women** inside the migraine subgroup.



(a) Men in total MIMIC population

(b) Men in migraine subgroup

Figure 4.14: Dendogram with a heatmap displaying the probability of a patient having two of the comorbid conditions associated to migraine, in this case for men in MIMIC-IV. A darker color codes for a lower p-value, thus a higher statistical relevance of the association between two conditions. In (a) we have the p-values associated to **men** in relation to the whole MIMIC population and their comorbid relations. In (b) we have the p-values for men's comorbidities correlations between **men** inside the migraine subgroup.

syndrome, Dystemic lupus erythematosus. The condition that seems to be most significantly associated to migraine male patients is gastroesophageal reflux disease and insomnia.

One obvious major difference that can be observed between women's and men's heatmaps is the intensity of the colors. In Figure 4.14 (b), we have lighter colors coding for a lesser relevant association of conditions, in the male population in the migraine subgroup, when compared to Figure 4.13 (b), for the female population. Some other relevant associations with conditions such as stroke and diabetes seem to have a higher likely hood of being associated to the most prominent conditions in men, while not being as relevant for women.

One important aspect to point out is that in all the heatmaps that were analyzed, it was possible to verify that in some areas of the heatmaps, there were some lighter "squares" of associations of comorbidities that seemed to have a higher p-value, and thus were considered to be not as relevant of a interaction. This means that, although this can been as these type of conditions do not interact with each other, this is not true. The meaning behind these is that a pair can still be highly correlated, and can appear in the totality of the population that was analyzed, meaning that because of that it is not as relevant. It does not stand out from other associations that are not seen as most commonly associated. This can be confirmed with the previous section of heatmaps associated to the counts of how many times these pairs of conditions occurred in the population.

4.3 Identification of subgroups within migraine patients in MIMIC-IV

One important way to analyse patients can be through dividing the population into subgroups through a certain common condition. This can be done through clustering, in order to see which features each group contains and characterize these groups. Clustering can upbring important subgroups of patients and give insight to attributes of these patients.

4.3.1 Clustering Analysis

Performing a clustering analysis to understand how patients with migraine are divided within this group is a practice that can be useful in characterizing this type of individuals. When given the patients' information such as demographics, comorbidities, transitions through the hospital for MIMIC-IV patients, the clustering model is able to retrieve which are the most important subgroups of the datasets depending on each of these features. For the totality of the 7 516 migraine patients and their relevant features, a clustering analysis was performed in an attempt to understand subgroups of this type of patients. In Table 4.2, we can assess these features in detail. The features that were taken into account are related to demographics of this group, such as the age and gender of patients, all the most common comorbidities related to migraine patient, which comprises in a total number of 34 conditions, and features related to the complexity of patients, such as the number of comorbidities each individual presents, the total number of ICD codes associated to each person and the number of hospital admissions.

However, since it could be concluded that some of the comorbidities have a prevalence of less than 0.01% in this migraine group, these conditions were removed of the features and the clustering was performed while eliminating them. Conditions such as chronic low back pain, periodontitis, parasomnia, advanced sleep phase, narcolepsy and sleep disordered breathing were removed from the features.

One crucial step when performing clustering is selecting which type of clustering method will be used, taking into account the context of which it is inserted into and the data at hand. It was decided to follow a similar approach to what is seen in Woldeamanuel et al. 2020, using a hierarchical clustering method with Gower's distance, and a PCA.

Table 4.2: Features extracted from dataset in order to go through the clustering process.

Type of features	Features
Demographics	Age, Gender
Comorbidities	Stroke, Diabetes, Obesity, Insulin resistance, Hypothyroidism, Endometriosis, Epilepsy, Major depressive disorder, Bipolar disorder, Post-traumatic stress disorder, Anxiety disorder, Fibromyalgia, Chronic low-back pain, Pain accompanying dysmenorrhea, Temporomandibular disorder, Insomnia Restless legs syndrome, Helicobacter pylori infection, Hepatobiliary disorders, Celiac disease, Irritable bowel syndrome, Inflammatory bowel disease, Constipation, Multiple sclerosis, Dystemic lupus erythematosus, Antiphospholipid syndrome, Primary Sjögren's syndrome, Rheumatoid arthritis, Atopic diseases
Complexity Parameters	Number of comorbidities, Total number of associated ICD-10 codes, Number of hospital admissions

Validating the clustering method that was used for both of the performances, it can be found in Table 4.3 the average SS computed for each of the *k* number of clusters, for the MIMIC-IV dataset.

Table 4.3: Values for the average sillhoute score and the corresponding number of *k* associated for the different methods of clustering for the MIMIC-IV dataset.

	MIMIC-IV	
	k	**Average Sillhoutte Score**
HC	2	0.417344
	3	0.355008
	4	0.149123
	5	0.126858
PCA+HC	2	0.423689
	3	0.342413
	4	0.165228
	5	0.13476

4.3.2 Hierarchical Clustering

Considering the validity methods that are necessary in order to perform a clustering analysis, the choice of the *k* clusters as to perform hierarchical clustering was based on a few factors. In Table 4.3, for the hierarchical clustering method without pairing it with PCA, the best SS is assigned to a number of clusters $k = 2$. However, when analysing the clusters population it was possible to find that for one of the clusters the total population had a small number of individuals, 380 compared to the 7 136 patients in the opposite cluster, and thus, did not demonstrate the desired outcome. Similarly to $k = 2$, the second best value was for $k = 3$, however, it divided the population in a similar way, having clusters with a low number of patients, and one big cluster with the majority of the individuals. Seeing these results, for the MIMIC-IV dataset, the number of 4 clusters was chosen as it demonstrated to have interesting results and represented the third highest value of the average SS for this method. Table 4.4 demonstrates which are the characteristics of the population of each the obtained clusters for the MIMIC-IV population. The first interesting result regarding this division of patients through $k = 4$ hierarchical clustering is that for the four groups, comparing the total number of patients, the cluster 3 contains the lowest value of 155 subjects, followed by cluster 4, which present 899 patients. The major differences between clusters features is seen through the median number of ICD codes, the median number of admissions and the number of comorbidities, as well as some of the comorbidities. There is heterogeneity among which of these conditions is high inside the clusters. However, anxiety disorder and gastroesophageal reflux disease, present a significant percentage among all four clusters. The ratio of female and male patients in cluster 1, 3 and 4 is steady and similar to the prevalence of the migraine in the population. For these three groups, females take up to 80% of the percentage of individuals in each cluster, while the 20% is related to the male population. Differently from cluster 2, which presents the highest percentage of female patients (87.36%) for a total number of 3 268 patients in this group. Alike what is happening to the female/male ratio, the average age for patients in cluster 1, 3 and 4 is similar, ranging from 49.42 to 58.42 years old. This can mean that the population in these subgroups are older than the ones in cluster 2, which presents a median age of 34.93.

Further analysing these subgroups, it is possible to understand that cluster 1 detains patients who have a median number of 2.67 comorbidities related to migraine, the lowest number of admissions. As these patients uptake the highest total number of individuals, this seems to be the majority of how migraine patients are represented. The most significant comorbidities related to this cluster are anxiety disorder (25.25%) and gastroesophageal reflux disease (41.67%). Other conditions that show up in at least 10% of this population are diabetes (17.92%), obesity (13.33%), hypothyroidism (15.26%), and constipation (11.39%). This cluster presents the oldest median age among all clusters.

The lowest number of ICD codes among the other clusters' patients belongs to cluster 2, with a number of 15.31, alongside with the lowest median number of comorbidities which are 2.42. This cluster presents the highest percentage of female patients of almost 90%, (87.36%), with the youngest median age of 34.93. The most significant disorder with a value of 31.49 of percentage is anxiety disorder, and conditions such as obesity (18.42%) and gastroesophageal reflux disease (17.04%) can be seen to have a somewhat percentage in this group. Looking at the most important differentiation that are age and the ratio of female to male patients, this group of patients can be related to the gender differences found in migraine patients. During women's fertile period of time, from ages of 20 to 50, women tend to have a higher prevalence of this disorder, as previously explored in Section 4.1. This falls into what is seen in this cluster, meaning that it comprises mostly female patients in their most fertile ages and can

Table 4.4: Characterization of the cluster population when performing hierarchical clustering and $k = 4$ clusters. Highlighted in bold the significant values with a threshold of 30% for percentages, and highlight high values of median numbers.

Features	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Total number of patients	3 204	3 268	155	899
Stroke (%)	4.28	3.06	7.74	7.31
Diabetes (%)	17.92	5.32	**52.9**	**42.07**
Obesity (%)	13.33	18.42	**50.97**	**44.66**
Insulin resistance (%)	0.25	0.06	1.94	0.67
Hypothyroidism (%)	15.26	9.3	**31.61**	**34.76**
Endometriosis (%)	0.25	3.3	5.81	2.92
Epilepsy (%)	3.9	9.33	**30.97**	21.48
Major depressive disorder (%)	1.47	3.55	20	9.56
Bipolar disorder (%)	3.03	6.88	25.81	18.22
Post-traumatic stress disorder (%)	2.75	7.56	28.39	22.38
Anxiety disorder (%)	25.25	**31.49**	**80**	**66.14**
Fibromyalgia (%)	3.18	1.1	9.68	14.17
Pain accompanying dysmenorrhea (%)	0.06	0.86	0.65	0.79
Temporomandibular disorder (%)	0.44	0.49	0.65	1.12
Insomnia (%)	8.3	5.02	**52.26**	**31.05**
Restless legs syndrome (%)	2.59	0.49	9.68	7.09
Gastroesophageal reflux disease (%)	**41.67**	17.04	**92.26**	**74.69**
Helicobacter pylori infection (%)	0.37	0.4	1.29	1.35
Hepatobiliary disorders (%)	0.28	0.21	3.23	1.46
Celiac disease (%)	0.75	0.92	1.29	1.91
Irritable bowel syndrome (%)	5.02	4.1	26.45	14.96
Inflammatory bowel disease (%)	2.75	2.75	10.97	7.2
Constipation (%)	11.39	6.67	**69.03**	**38.58**
Multiple sclerosis (%)	0.91	1.53	2.58	1.57
Dystemic lupus erythematosus (%)	0.78	1.71	3.87	3.15
Antiphospholipid syndrome (%)	0.16	0.31	0.65	0.79
Primary Sjögren syndrome (%)	0.66	0.46	4.52	1.69
Rheumatoid arthritis (%)	0.06	0.03	0.65	0.11
Atopic diseases (%)	0.09	0.06	0.65	0.11
Female (%)	**73.22**	**87.36**	**70.32**	**76.94**
Male (%)	26.78	12.64	29.68	23.06
Median age (years)	**58.63**	34.93	49.42	53.42
Median number of comorbidities	2.67	2.42	**7.26**	5.72
Median number of ICD codes	18.93	15.31	**153.92**	64.59
Median number of admissions	2.41	2.53	**36.24**	9.67

be related to the hypothesis that migraine is highly associated to sex hormones which are key in this period of time.

The group of migraine patients that express a higher variability can be seen in cluster 3, where there is a high number of a wide variety of features. This group englobes individuals who were identified to have the highest number of ICD codes associated, with a median number of 153.92 different codes associated, the highest median number of 36.24 admissions and the highest number of 7.26 median comorbidities related to migraine. Thus, it is plausible to speculate that this group portrays patients associated to multiple conditions, relating this group to multimorbidity patients. These type of patients are highly complex and it can be verified through the characteristics presented within this cluster. With a median age of 49.42, these individuals present an elevated percentage within all the set of comorbid conditions. The highest percentage is linked to gastroesophageal reflux disease, which is associated to 92.26% of the population in this group. The following highest numbers are related to anxiety (80%), constipation (69.03%), and conditions such as diabetes (52.9%), obesity (50.97%) and insomnia (52.26%) appear in at least half of the population.

For cluster 4, it is possible to assess that this cluster has more similarities to cluster 3. This group has the oldest median age for individuals and apart from demonstrating some of the same comorbidities among the clusters, it contains significant higher values than cluster 1 and 2 for conditions such as diabetes (42.07%), obesity (44.66%), hypothyroidism (34.76%), anxiety disorder (66.14%), insomnia (31.05%), gastroesophageal reflux disease (74.69%) and constipation (38.58%). With a median number of 64.59 ICD codes, 5.72 comorbidities and 9.67 admissions, this cluster can be seen as a combination between the cluster 1 and cluster 3, taking some of the most common characteristics between both clusters, comprising of a total of 899 patients.

Overall, the factor that has divided these migraine patients into their designated groups has been identified as being related to the median number of total ICD codes, the median number of admissions and comorbidities, as well as some striking comorbidities such as diabetes, obesity, insomnia, and constipation. One important dividing feature that can also be seen is gender, as it is an outstanding feature for cluster 2.

Comparing these results to what was previously seen through the preceding Sections, it confirmed that some of the comorbidities such as gastroesophageal reflux disease, obesity, diabetes, insomnia and constipation are highly related to migraine patients. It was possible to also observe some of the gender differences among this type of individuals, as one of the clusters contained a percentage of almost 90% women with ages ranging from 20-50 years old.

4.3.3 Principal Components Analysis and Hierarchical Clustering

Performing a PCA can be beneficial when there is a big amount of information. As seen in Section 2.5.4, some of the studies that involved migraine patients and clustering methods have opted for doing a PCA analysis in order to attain the results in an easier manner. Thus, it was one of the steps taken in this dissertations' clustering process. After performing the hierarchical clustering method by itself, the principal components of the features were found. In order to do this, it was resorted to the Python's machine learning library, scikit-learn. Because PCA aims to reduce the dimensionality of the data, using the *fit_transform* function of the already standardized data of migraine patients, with all the necessary features for clustering, it was possible to obtain the principal components. Similarly to what is requested for performing a hierarchical clustering analysis, during PCA the information must be unbiased, and contain a variance of similar values in all the different features. One important aspect that should be analyzed when performing the PCA is the proportion of the variance of each of the principal components, meaning which of the features contributes the most to this variance. The first principal components usually detain the higher values of variances, so the order at which these components are presented has an impact, and they can be correlated to the original features in the dataset. This can be seen in Figure 4.15, as it is possible to verify which of the features has contributed the most for each of the principal components. In order to analyse this correlation matrix, it must be taken into account that the features that present values near zero, and thus have no contribution to explaining the principal component are represented in yellow tones, having a neutral role. On the other hand, for the maximal and minimal values that are represented in blue and red tones respectively, the darker its tone, the more important this feature is as to explain the principal role, and thus contributing more to it. Since the order at which the principal components are presented is of importance, one must analyse in detail which features contribute the most to the variation of these. Looking into the first principal components, it is possible to assess that some features such as the total number of comorbidities, the total number of ICDs per patient, and the number of admissions play a crucial role in the first principal component. In the second component, age is deemed as an important factor and presents a high negative value. Gender on the other hand has a neutral role in the first two principal components, along with the majority of the comorbidities. However, starting from principal component number three, up until the eleventh principal component, there is some variation and thus a contribution of this feature. Some of the most important comorbidities for the initial principal components are diabetes, obesity, anxiety disorder, and gastroesophageal reflux disease. This confirms what has already been seen in the previous Section, as these were one of the distinctions between the clusters and the conditions that were shown in the subgroups of migraine patients. To understand how each of the principal components can explain the data and its variance, it can be seen in Table 4.5 the values of the explained variation, alongside the cumulative proportion of the whole dataset. In the MIMIC-IV dataset, when performing a PCA analysis, and considering the optimal window of 70-95% of the coverage of the variation, only the twelve first principal components were taken into account. These components amount for a percentage of 91% of the variation, and adding more components from that point on, only accounts for a small percentage of change in the cumulative variation, as seen in Table 4.5. Thus, the selection of the first twelve components as the data for the performance of the hierarchical clustering, as a form of reducing the complexity of the information. Nevertheless, principal components are a way to reduce dimensionality, and there is no real interpretation behind each of them. They comprise information of all the features together while retaining as much information as possible from the original dataset, but they can not be analyzed by themselves.

As a validating feature, migraine was taken into account and it can be seen that it contributed for the last principal component, as all the patients had this condition, and thus, the variance was not of importance. This can be seen in Figure 4.15.

By choosing the first twelve principal components of the dataset, it was possible to perform hierarchical clustering by resorting to the reduced data. As it can be verified in Table 4.5, principal component

Table 4.5: Variance that can be explained for each of the principal components, as well as their cumulative proportion. PC - Principal Component

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC10	PC11	PC12	PC13	PC14	PC15	...	PC37
Variance	$4.4501e^{-1}$	$1.7572e^{-1}$	$1.0402e^{-1}$	$3.2287e^{-2}$	$2.8531e^{-2}$	$2.8282e^{-2}$	$2.3154e^{-2}$	$2.1714e^{-2}$	$1.7938e^{-2}$	$1.5810e^{-2}$	$1.3953e^{-2}$	$1.1404e^{-2}$	$1.0333e^{-2}$	$8.6649e^{-3}$	...	$1.8343e^{-33}$
Cumulative variance	0.4450	0.6207	0.7247	0.7570	0.7855	0.8138	0.8370	0.8587	0.8952	0.9110	0.9249	0.9363	0.9467	0.9553	...	1

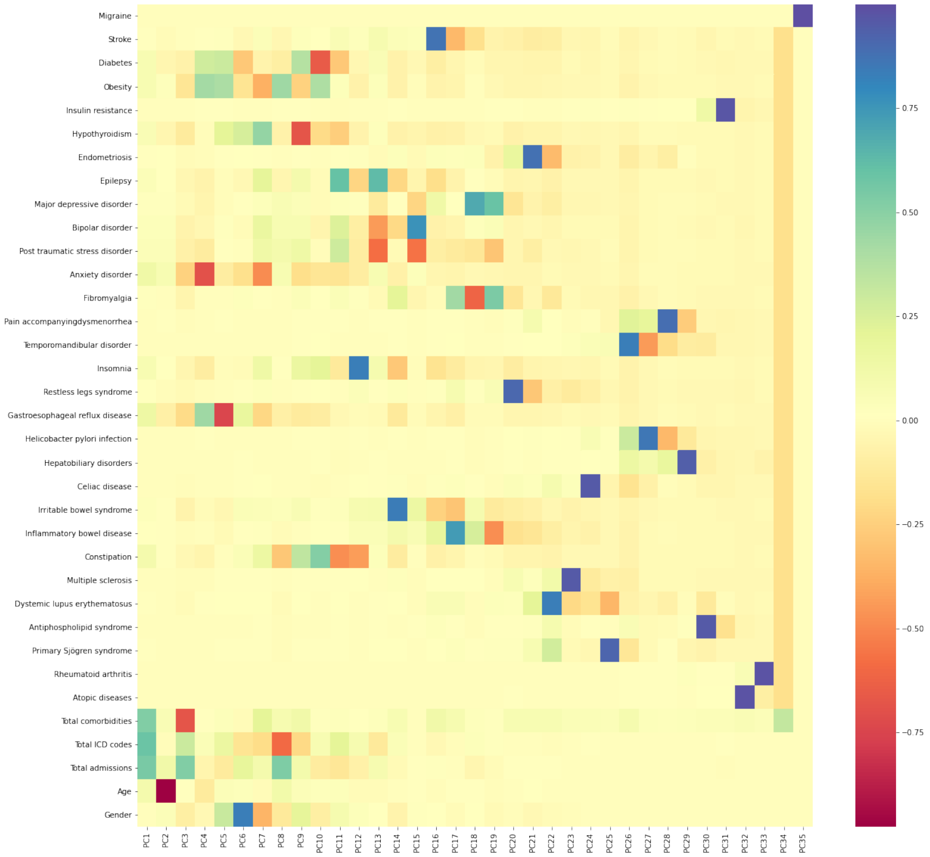


Figure 4.15: Generated correlation matrix plot between the features used for clustering and the principal components. Higher values are coded as blue, neutral values are yellow and lower values are represented in red. PCX - Principal Component, where X is the number of said component.

number 12 can explain 92% of the variance of the data, and as the number of principal components gets higher, the percentage at which they explain is steadily lower. Thus, the choice of using only 12 of the principal components was taken as a reasonable one.

Assessing the best average SSs and their relationship to the *k* clusters, it was decided to assign four clusters, although it did not perform as the highest SS, but showed to have more compact results. While exploring the clusters, the data showed to contain a clear differentiation between the four groups. Similarly to what has happened before performing this PCA analysis, the percentage of female and male follows the trend found in literature and what has been seen in Section 4.3.2, where women portray 80% of the individuals in each cluster, while men take up to 20% for cluster 1, cluster 3 and 4.

Similar to what was seen in hierarchical clustering by itself, the most common comorbidity that assumes a high percentage among all of the four clusters is anxiety disorder. The differentiation factor of the clusters can be observed through the median number of ICD codes, admissions and comorbidities, mirroring what could be observed through the first hierarchical analysis. This also confirms what was seen through the analysis of the correlation between features, as these were the three features that contributed the most to the variance of the original data and explain the first principal component seen in Figure 4.15.

It is possible to see the characteristics of the clusters in Table 4.6. In cluster 1, the total number of patients is the highest, comprising 3 972 patients. The percentages related to the comorbidities are all below 50%, and in a smaller number of conditions comparing to cluster 3 and 4. Anxiety disorder (27.84%) and gastroesophageal reflux disease (35.35%) present the highest percentage among this group. With a number of 2.65 median comorbid conditions and 18.55 number of ICD codes, these patients have a median number of admissions of 2.44.

Analyzing the characteristics of patients in cluster 2, these contain 2 442 individuals and it is similar to cluster 2 of the HC analysis. Female migraine patients who express a median of two of the comorbidities, with a lower number of ICD codes and admissions. For the total number of patients, a percentage of 89.56% is women, similarly to what was seen in cluster 2 of HC analysis. Having the youngest median age among the other clusters, this cluster comprises information about mostly female patients in their fertile period, as previously explained. With the highest percentage of anxiety disoder (27.03%) as the condition that affects the most among this group of migraine patients.

As for cluster 3 and 4, it is possible to assess that these clusters contain similarities by grouping individuals who have a high spectrum of different conditions, having both higher number of comorbidities, ICD codes and admissions when comparing to clusters 1 and 2. Although the number of patients for both clusters is low, the complexity of these patients can be seen through the percentage at which they express the features. In cluster 3, conditions such as diabetes (36.8%), obesity (45.03%), hypothyroidism (33.26%), post-traumatic stress disorder (20.11%), anxiety disorder (68.34%), insomnia (31.89%), constipation (40.46%) are also seen in cluster 4, and with the highest percentage in both of them can be seen gastroesphageal reflux disease with 77.26% in cluster 3 and 91.19% in cluster 4. One condition that has a higher percentage in cluster 4 is epilepsy, with 30.84%, comparing to a 18.4% in cluster 3. The main differentiation of these clusters lays in the total number of median comorbidities, with cluster 4 being more complex patients who present a median of 7.19 conditions related to comorbidity, and a median number of 139.04 ICD codes, with the highest median number of 31.04 admission. Thus shows how complex the patients in cluster 4 are, being associated to multiple diseases, and not only comorbidities related to migraine. This division through PCA has allowed to confirm the previous

Table 4.6: Characterization of the cluster population when performing principal component analysis and hierarchical clustering for $k = 4$ clusters of the MIMIC-IV population. Highlighted in bold the significant values with a threshold of 20% for percentages, and highlight high values of median numbers.

Features	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Total number of patients	3 972	2 442	875	227
Stroke (%)	4.68	2.17	6.17	9.25
Diabetes (%)	17.85	2.38	**36.8**	**50.66**
Obesity (%)	14.88	16.67	**45.03**	**49.78**
Insulin resistance (%)	0.23	0.04	0.57	1.76
Hypothyroidism (%)	14.48	8.35	**33.26**	**35.68**
Endometriosis (%)	1.03	2.99	2.97	4.85
Epilepsy (%)	6.14	7.94	18.4	**30.84**
Major depressive disorder (%)	1.64	4.05	8.11	19.38
Bipolar disorder (%)	3.4	8.11	15.09	**25.99**
Post-traumatic stress disorder (%)	3.58	8.15	**20.11**	**26.87**
Anxiety disorder (%)	**27.84**	**27.03**	**68.34**	**81.94**
Fibromyalgia (%)	2.64	1.68	12.57	10.13
Pain accompanying dysmenorrhea (%)	0.08	1.06	0.8	0.88
Temporomandibular disorder (%)	0.45	0.41	1.26	0.88
Insomnia (%)	7.63	3.56	**31.89**	**51.98**
Restless legs syndrome (%)	2.09	0.49	6.97	9.25
Gastroesophageal reflux disease (%)	**35.35**	16.87	**77.26**	**91.19**
Helicobacter pylori infection (%)	0.38	0.37	1.03	2.64
Hepatobiliary disorders (%)	0.23	0.2	1.6	2.64
Celiac disease (%)	0.76	1.02	1.71	1.32
Irritable bowel syndrome (%)	4.41	5.16	14.06	19.82
Inflammatory bowel disease (%)	2.59	2.66	7.54	11.01
Constipation (%)	9.77	5.73	**40.46**	**66.52**
Multiple sclerosis (%)	1.06	1.56	1.37	2.2
Dystemic lupus erythematosus (%)	1.11	1.31	3.09	5.29
Antiphospholipid syndrome (%)	0.2	0.29	0.57	1.32
Primary Sjögren syndrome (%)	0.63	0.25	2.06	3.96
Rheumatoid arthritis (%)	0.05	0	0.23	0.44
Atopic diseases (%)	0.05	0.08	0.23	0.44
Female (%)	**74.55**	**89.56**	**77.94**	**72.25**
Male (%)	**25.45**	10.44	**22.06**	**27.75**
Median age (years)	**56.19**	31.56	52.41	48.56
Median number of comorbidities	2.65	2.31	5.6	**7.19**
Median number of ICD codes	18.55	14.31	59.18	**139.04**
Median number of admissions	2.44	2.43	8.64	**31.04**

obtained results in Section 4.3.2. However, the limiting factor for this analysis is that it explains 91% of the data, as only the first twelve principal components were taken for the hierarchical clustering. Thus, one may look carefully at these results.

4.4 eICU-CRD analysis

Through the initial analysis on the information found about migraine patients in the eICU-CRD dataset, it was decided to perform a simpler analysis regarding these individuals. The reasons that lead us to simplify the analysis was the reduced number of patients that presented this condition and the context at which these patients are inserted, which is critical care. Nevertheless, it is important to assess how the comorbidities occurrences can occur within this population, as a complement to the already studied analysis of the MIMIC-IV dataset.

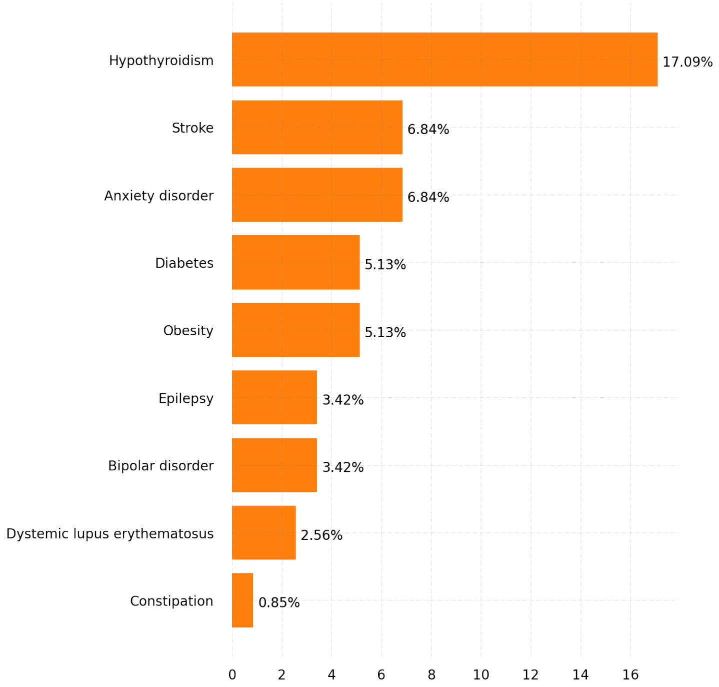
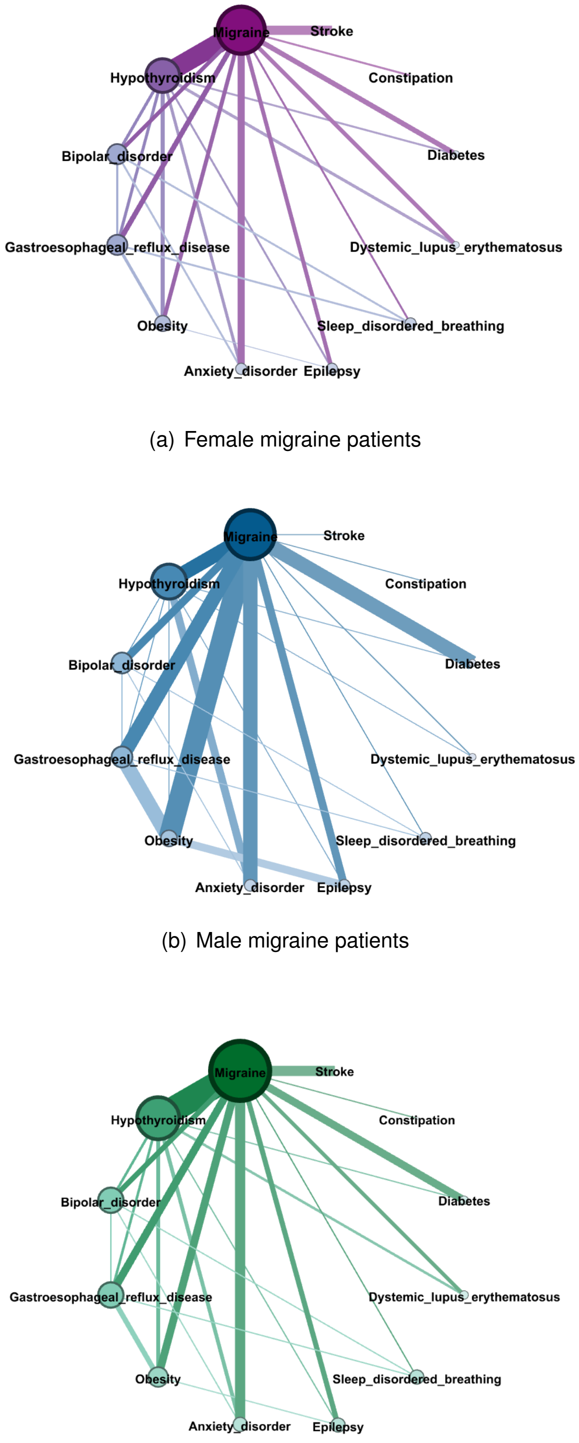


Figure 4.16: Comorbidities related to migraine and their percentage in the subgroup of patients that are affected by this condition.

Similar to what has been done for the MIMIC-IV dataset, a network analysis for the total population of migraine in the eICU-CRD dataset, as well as the division between both genders was performed. The obtained results can be seen through Figure 4.17. Comparing these results to what was seen in MIMIC-IV, it is possible to verify that this group of migraine patients presents a smaller number of associated comorbidities, making it easier to analyze. Since the extreme conditions of patients in critical care units are difficult to manage, and often do not allow for a deeper understanding of what the patient is feeling, some conditions such as insulin resistance, endometriosis, major depressive episode, post-traumatic stress disorder, fibromyaldiga, chronic low back pain, pain accompanying dysmenorrhea, temporomandibular disorder, insomnia, restless legs syndrome, helicobacter pylori infection, hepatobiliary disorders, celiac disease, irritable bowel syndrome, inflammatory bowel disease, multiple sclerosis, antiphospholipid syndrome, primary sjogren syndrome, rheumatoid arthritits, atopic diseases were not taken as a diagnosis for these patients, and are not a part of the patients' diagnosis. Thus not shown in these graphs or in the following results. The three networks represented in Figure 4.17 show a similarity between each other, with migraine being the biggest node in size of the three networks, and is highly correlated to conditions such as hypothyroidism. The main difference found between these networks is the weight that each edge is associated to.



(a) Female migraine patients

(b) Male migraine patients

(c) Total migraine patients

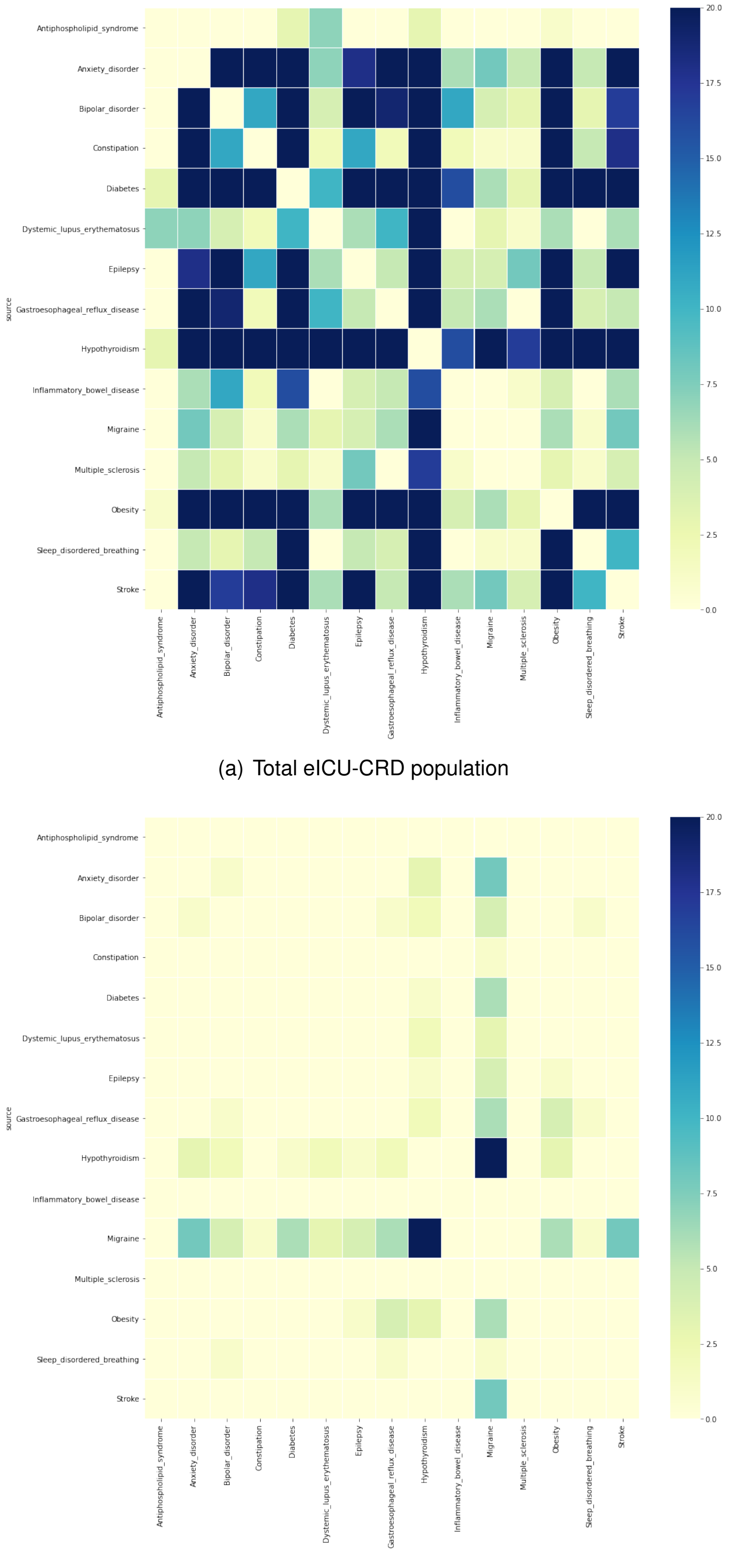
Figure 4.17: Circular graphs represent how the most common comorbidities of migraine are related to each other. Nodes represent the disorders, and the edges represent the times each combination of two occurs. (a), (b) and (c) have 12 nodes and 24 edges.

Trying to understand how the found comorbidities within this dataset are connected to each other, in a similar way it was computed the heatmaps for the totality of the migraine patients. In Figure 4.18, it is possible to see which conditions are most commonly seen through patients within both the totality of eICU-CRD patients in (a) and the migraine patients of this dataset (b). It is important to take into account the fact that this dataset contains patients who are in a critical care context, meaning that some conditions were not reflected in the heatmaps, when compared to the other dataset. The top value of occurrences for each pair of conditions was given to be 200, meaning that darker shades account for a higher value of counts among the totality of the pairs. In (a), it is possible to see that these conditions are highly seen together in the eICU-CRD population, and hypotyroidism is highly related to all the comorbidity conditions of migraine present in the dataset, except for antiphospholipid disorder. When it comes to the gender analysis, it can be seen through Figure 4.19 that the totality of female patients for eICU-CRD has a similar heatmap to what is seen for the totality of eICU-CRD patients in Figure 4.18 (a). However, men display different outcomes, for both the total population and the migraine-related male population. This can be seen in Figure 4.20, where in (b), the total number of counts for the pairs of conditions is seen to be below 25.

In Figure 4.21, we can see how relevant the association between two conditions for the totality of the migraine patients. In this case, it is possible to see through the dendogram that the association between migraine with gastroesophageal reflux disease and multiple sclerosis, and antiphospholipid syndrome. For the majority of the comorbid conditions present in this population, pairs that contain the migraine condition are associated to the lowest values of p-value, except for conditions such as inflammatory bowel disease, antiphospholipid disorder and multiple sclerosis, which have a higher p-value, meaning its relevance to the migraine patients is less significant. Nevertheless, they seem to be the closest ones connected to the migraine condition.

Dividing the migraine population into genders, it is possible to assess the heatmaps associated to dendograms in Figure 4.21 (a) for total women and (b) for migraine female population, how the comorbidities in these groups are associated to each other. Similarly to what is shown in Figure 4.21, when diving the population through gender it is possible to find some of the comorbidities that are mostly related to migraine and to each other. In the totality of women, the conditions seem to have a similar relationship to what is found for the total eICU-CRD population. Contrarily to what is seen in the migraine subgroup of female patients, there is a high number of associations among disease that has a low p-value, meaning that these conditions are significant when paired together within this population. This can be seen in Figure 4.22. Comparing these results to what is seen in the MIMIC-IV population, the distribution of the low p-value within the pairs of conditions among female patients that have migraine is similar. There are a greater number of pairs of conditions that are relevant to this group and that is seen through the darker colors within both Figure 4.13 (b) and Figure 4.22 (b). Regarding the male population, it is possible to verify that for the totality of male patients in Figure 4.23 (a), and for migraine male patients in (b), there is some consistency in groups of disorders that do not seem to have a significant correlation. Namely, dystemic lupus, inflammatory bowel disease and sleep-disordered breathing, antiphospholipid syndrome, migraine and gastroesophageal disease, with multiple sclerosis.

As for the network analysis of eICU-CRD's statistically relevant networks, the conditions which seem to be highly connected to migraine are hypothyroidism, obesity, stroke, diabetes and anxiety disorder.

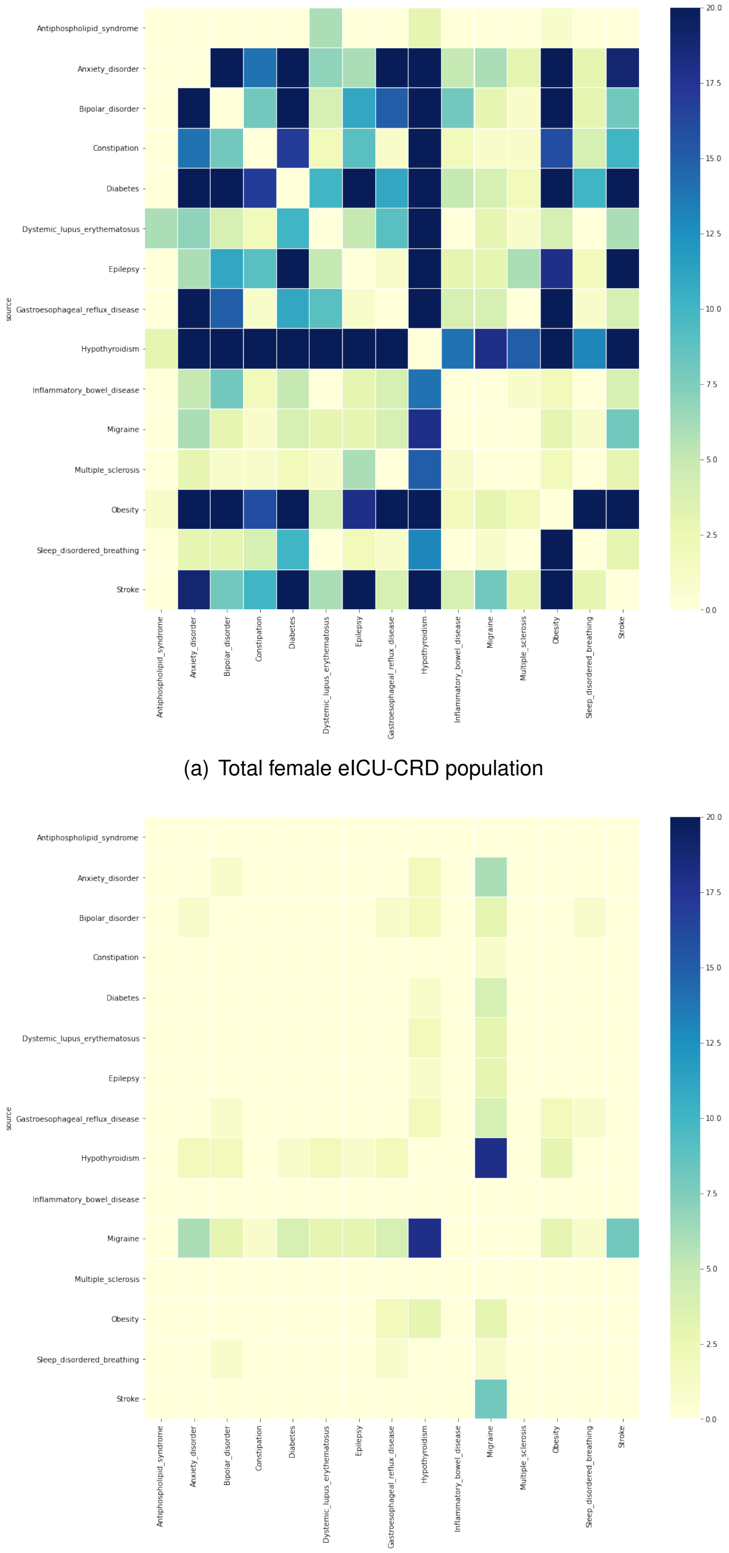


(a) Total eICU-CRD population

(b) Total migraine population

Figure 4.18: Heatmap displaying the counts of two of the comorbid conditions associated to migraine, in this case for the whole population and the migraine subgroup. Darker shade of blue codes for a higher appearence, opposed to lighter shades of blue/white, coding for lowest values.

This can be verified in Figure 4.24 for the totality of the eICU-CRD patients. This confirms what was already seen through the MIMIC-IV analysis, as these comorbid conditions are very much connected

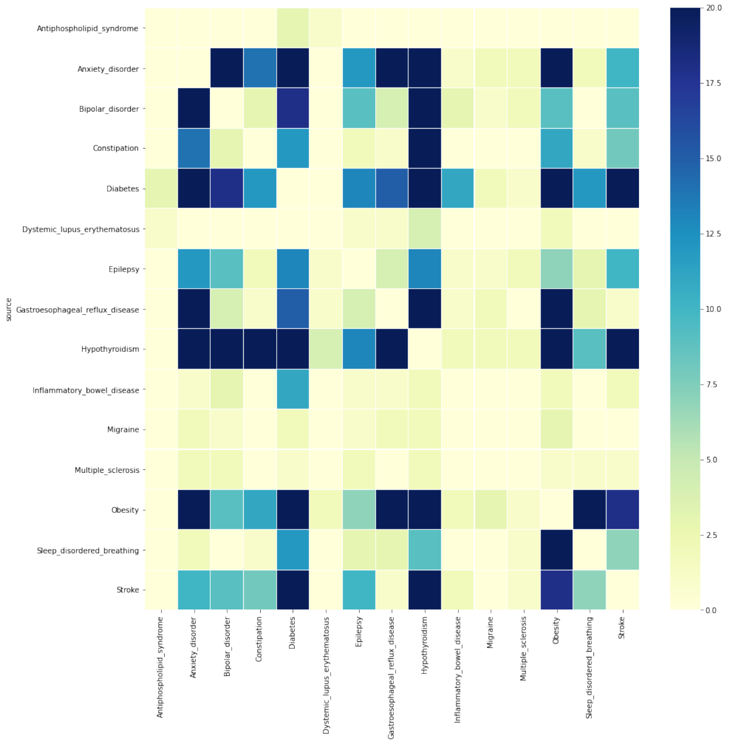


(a) Total female eICU-CRD population

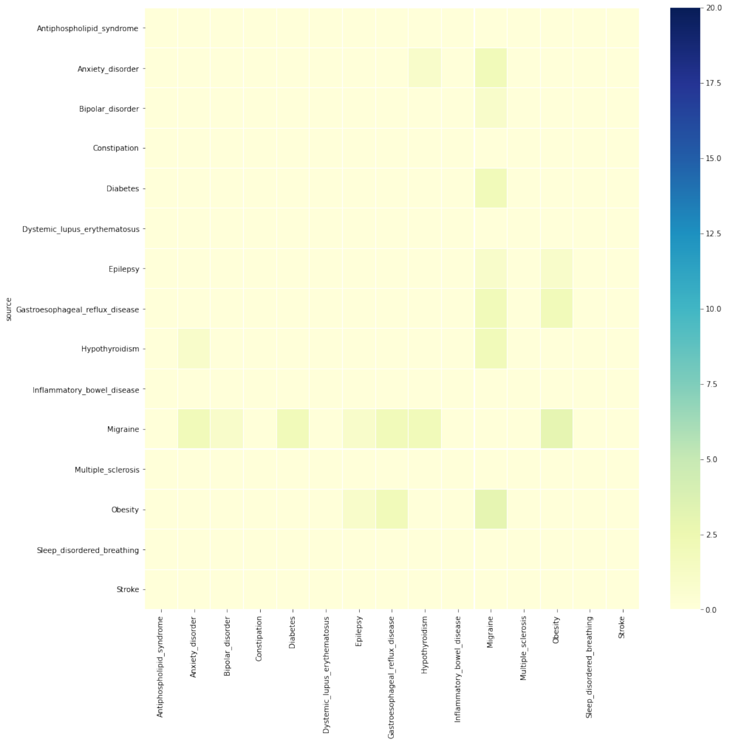
(b) Total female migraine population

Figure 4.19: Heatmap displaying the counts of two of the comorbid conditions associated to migraine, in this case for the whole female population and the female migraine subgroup. Darker shade of blue codes for a higher appearence, opposed to lighter shades of blue/white, coding for lowest values.

to each other and to migraine. The graphs for the female total eICU-CRD and migraine population can be found in Figure 4.25 (a) and (b), respectively. In this case, for the female migraine population,



(a) Total male MIMIC population



(b) Total male migraine population

Figure 4.20: Heatmap displaying the counts of two of the comorbid conditions associated to migraine, in this case for the whole male population and the male migraine subgroup. Darker shade of blue codes for a higher appearence, opposed to lighter shades of blue/white, coding for lowest values.

stroke seems to be highly connected to migraine. Conditions such as hypothyroidism , obesity, gastoesophageal reflux disease and bipolar disorder are shown to have connections between each other more

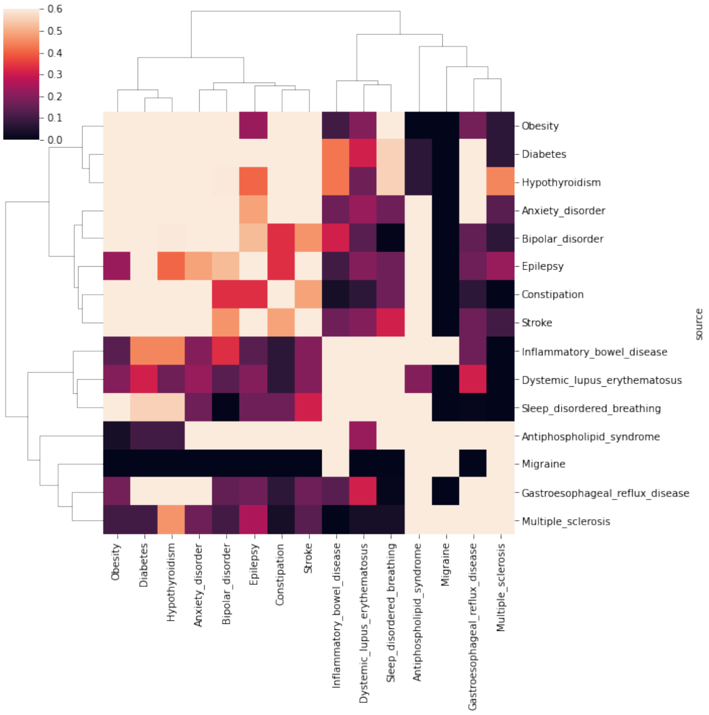
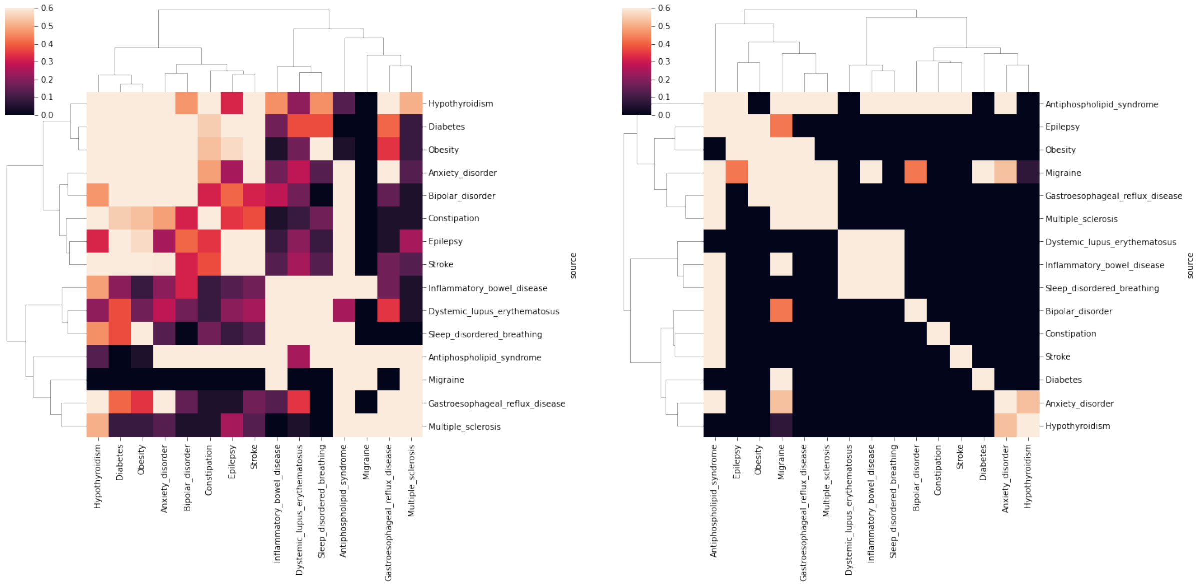


Figure 4.21: Dendogram with a heatmap of the total migraine population and displaying the probability of a patient having two of the comorbid conditions associated to migraine. A darker color codes for a lower p-value, thus a higher statistical relevance of the association between two conditions.

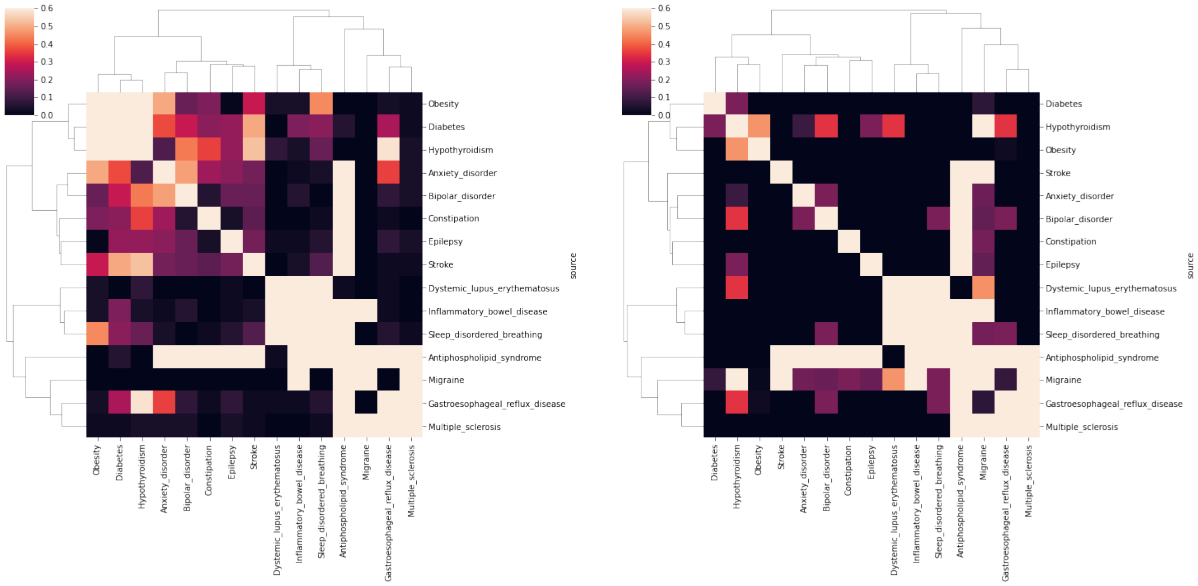


(a) Women in total eICU-CRD population

(b) Women in migraine subgroup

Figure 4.22: Dendogram with a heatmap displaying the probability of a patient having two of the comorbid conditions associated to migraine, in this case for women. A darker color codes for a lower p-value, thus a higher statistical relevance of the association between two conditions. In (a) we have the p-values associated to women in relation to the whole eICU-CRD population and their comorbid relations. In (b) we have the p-values for women's comorbidities correlations between women inside the migraine subgroup.

strongly than to migraine. As for the male population, in Figure 4.26, it is possible to see that there are



(a) Men in total eICU-CRD population

(b) Men in migraine subgroup

Figure 4.23: Dendogram with a heatmap displaying the probability of a patient having two of the comorbid conditions associated to migraine, in this case for men. A darker color codes for a lower p-value, thus a higher statistical relevance of the association between two conditions. In (a) we have the p-values associated to women in relation to the whole eICU-CRD population and their comorbid relations. In (b) we have the p-values for women's comorbidities correlations between men inside the migraine subgroup.

obvious differences for the total male population of eICU-CRD and the migraine male population. Men who experience migraine are highly associated to obesity, with gastroesophageal reflux disease and epilepsy as the combinations of disorders that are also significant for this type of patient.

For the clustering analysis, it was decided to only perform the Hierarchical Clustering method, without resorting to PCA, due to the fact that the number of patients in this dataset is reduced, as well as the number of features. Features that were not used in this clustering: total number of admissions, insulin resistance, endometriosis, major depressive episode, post-traumatic stress disorder, fibromyaldiga, chronic low back pain, pain accompanying dysmenorrhea, temporomandibular disorder, insomnia, restless legs syndrome, helicobacter pylori infection, hepatobiliary disorders, celiac disease, irritable bowel syndrome, inflammatory bowel disease, multiple sclerosis, antiphospholipid syndrome, primary sjogren syndrome, rheumatoid arthritis, atopic diseases. Meaning that a small number of conditions were analyzed, simplifying the clustering performance. Since the number of patients with migraine in eICU-CRD is reduced, this reflects on the performance of the clustering method. In Table 4.7 it is possible to see the values the SS values that were computed, and $k = 4$ was chosen. This allowed to see 4 different groups of patients for which the characteristics can be evaluated through Table 4.8.

Cluster 1 comprises patients who are only associated to migraine and have none of the comorbid conditions related to this disorder found in the eICU-CRD dataset. These patients have a reduced number of ICD codes associated to them. For cluster 2 and cluster 4, it can be seen that the ratio of female to male patients is similar, and these two formed groups have a high median number of ICD codes. For cluster 3, it was possible to verify a group of 90% female patients, with a median age of 32.62. Similar to what was found in MIMIC-IV, this cluster corresponds to what is seen in the literature

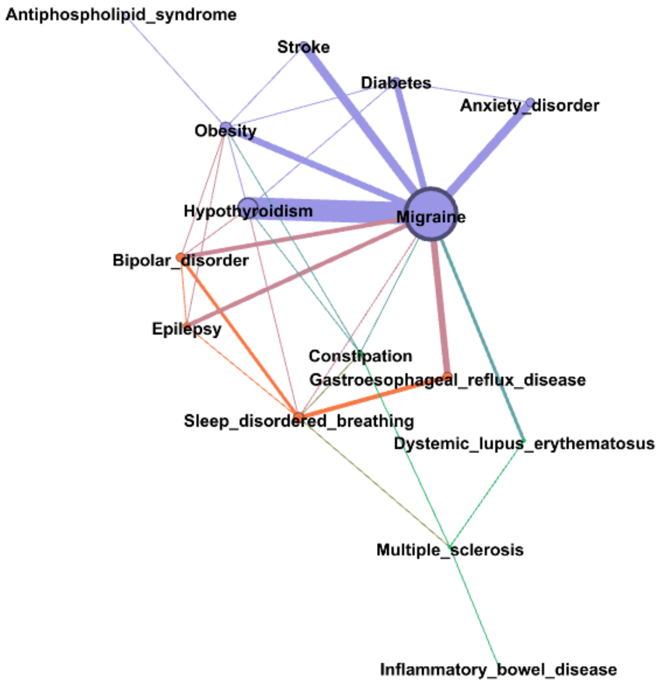
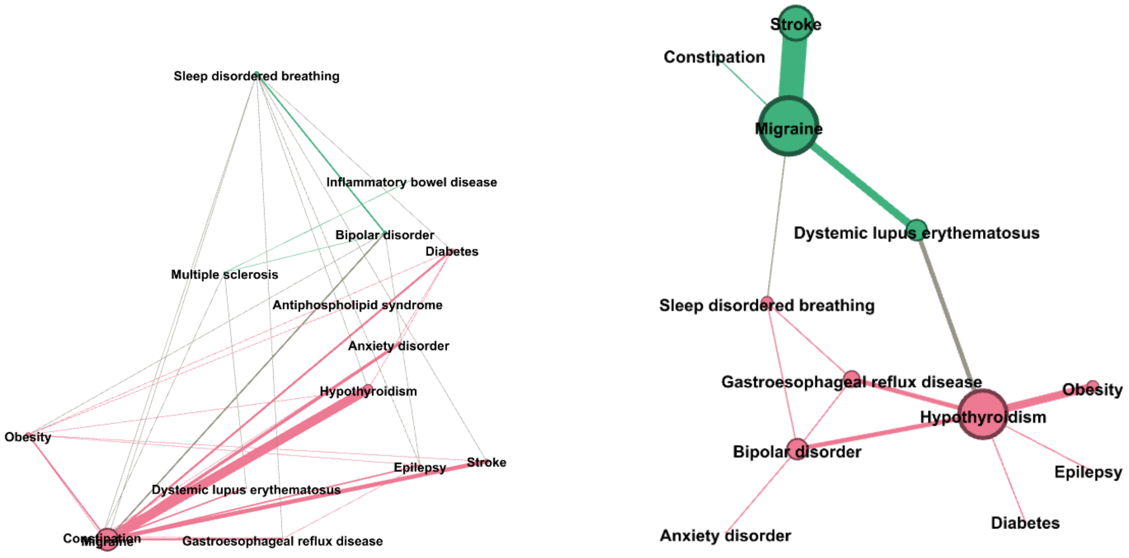

Antiphospholipid_syndrome
Stroke
Diabetes
Anxiety_disorder
Obesity
Hypothyroidism
Migraine
Bipolar_disorder
Epilepsy
Constipation
Gastroesophageal_reflux_disease
Sleep_disordered_breathing
Dystemic_lupus_erythematosus
Multiple_sclerosis
Inflammatory_bowel_disease

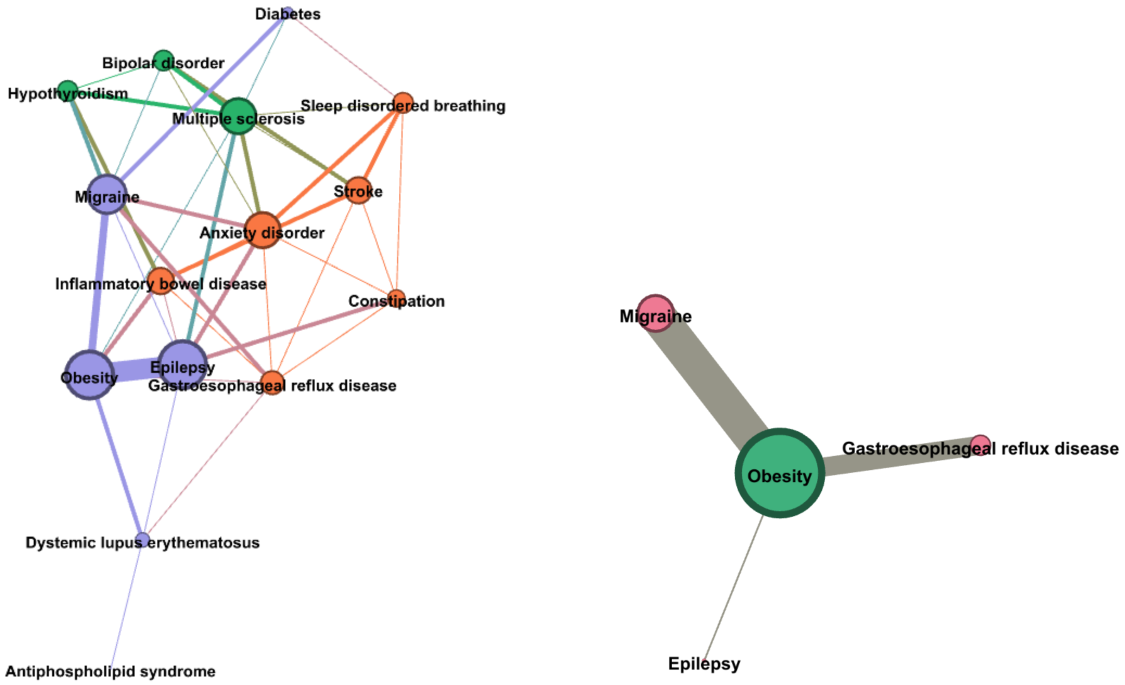
Figure 4.24: Representation of the graphs containing the most relevant combinations of comorbididities within the **total** population of the eICU-CRD dataset and the total migraine population, which stated a p-value < 0.05. Contains 15 nodes, 32 edges.


Sleep disordered breathing
Inflammatory bowel disease
Bipolar disorder
Diabetes
Multiple sclerosis
Antiphospholipid syndrome
Anxiety disorder
Hypothyroidism
Obesity
Epilepsy
Stroke
Dystemic lupus erythematosus
Constipation
Migraine
Gastroesophageal reflux disease
Stroke
Constipation
Migraine
Dystemic lupus erythematosus
Sleep disordered breathing
Gastroesophageal reflux disease
Obesity
Hypothyroidism
Bipolar disorder
Epilepsy
Diabetes
Anxiety disorder

(a) Women in total eICU-CRD population

(b) Women in migraine subgroup

Figure 4.25: Representation of the graphs containing the most relevant combinations of comorbididities within the **female** population eICU-CRD dataset and the migraine female population, which stated a p-value < 0.05. (a) Contains 15 nodes, 36 edges. (b) Contains 12 nodes, 14 edges.


Diabetes
Bipolar disorder
Hypothyroidism
Multiple sclerosis
Sleep disordered breathing
Stroke
Migraine
Anxiety disorder
Inflammatory bowel disease
Constipation
Obesity
Epilepsy
Gastroesophageal reflux disease
Dystemic lupus erythematosus
Antiphospholipid syndrome
Migraine
Gastroesophageal reflux disease
Obesity
Epilepsy

(a) Men in total eICU-CRD population

(b) Men in migraine subgroup

Figure 4.26: Representation of the graphs containing the most relevant combinations of comorbididities within the **male** population eICU-CRD dataset and the migraine male population, which stated a p-value < 0.05. (a) Contains 15 nodes, 41 edges. (b) Contains 4 nodes, 3 edges.

Table 4.7: Values for the average SS and the corresponding number of *k* associated for the different methods of clustering for the MIMIC-IV dataset.

	eICU-CRD	
	k	**Average Sillhoutte Score**
HC	2	0.277073
	3	0.227722
	4	0.24034
	5	0.244837

about women in their childbearing ages. It is also interesting to point out that this Cluster contains the highest value of patients associated to stroke, with a percentage of 21.88. Connecting this to the risk factor of cardiovascular disorders seen in female migraine patients, this cluster seems to correspond to what is seen in the literature. Cluster 2 contains the highest comorbid conditions within all clusters, with the highest percentages for the comorbid conditions and cluster 4 is associated to patients with stroke, bipolar disorder, and anxiety disorder.

Table 4.8: Characterization of the cluster population when performing principal component analysis and hierarchical clustering for $k = 4$ clusters. Highlighted in bold are the significant values with a threshold of 20% for percentages, and highlight highest values of median numbers.

Features	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Total number of patients	47	26	32	12
Stroke (%)	0	0	**21.88**	8.33
Diabetes (%)	0	7.69	12.5	0
Obesity (%)	0	19.32	3.12	0
Hypothyroidism (%)	0	**73.08**	3.12	0
Epilepsy (%)	0	3.85	9.38	0
Bipolar disorder (%)	0	11.54	0	8.33
Anxiety disorder (%)	0	19.23	3.12	16.67
Constipation (%)	0	3.85	0	0
Dystemic lupus erythematosus (%)	0	7.69	3.12	0
Female (%)	**70.21**	**84.62**	**90.62**	**83.33**
Male (%)	**29.79**	15.38	9.38	16.67
Median age (years)	**58.68**	58.54	32.62	46.17
Median number of comorbidities	1.02	**2.69**	1.56	1.33
Median number of ICD codes	4.51	9.19	3.78	**12**

Chapter 5

Conclusions

The presented work aims to characterize patients who are associated with the migraine condition and understand how their most common comorbid disorders are associated with each other while resorting to network science and clustering methods. Applying these methods to two different datasets, one of hospitals' wide EHR which is MIMIC-IV and another focused on critical care unit patients (eICU-CRD), it was possible to understand how the most common comorbidities associated with migraine are related to each other, and how the individuals who suffer from migraine are grouped together based on specific characteristics. Migraine patients have a wide spectrum of conditions that are associated with it, and several already known gender differences among patients. Literature has shown that it can be beneficial to explore these differences and understand better the migraine population, as well as how these conditions are related to each other. The obtained conclusions from this dissertation are presented in this Chapter, followed by the limitations that it had, as well as what could be the possible future work for the proposed work related to migraine patients.

5.1 Achievements

The data of the patients from the two datasets were retrieved in order to understand the population as a whole and to see which characteristics were presented among each of the databases. The data followed a pre-processing stage, in which there was an initial step needed to convert the diagnosis codes of patients from 9$^{\text{th}}$ to 10$^{\text{th}}$ revisions of the ICD system. After this, the next step was to select only patients who presented either the condition of migraine or the most common comorbid disorders that have already been pre-identified. After the pre-processing stage, migraine patients were characterized for both databases, and the analyses performed were done in order to understand the correlation between migraine comorbid diseases using network science, and to uncover which groups of patients were formed within this group of individuals, based on features and through clustering methods.

Using networks in order to visualize and further understand how the comorbid conditions are interconnected among migraine patients, it was possible to verify some gender differences. Conditions such as anxiety disorder, gastroesophageal reflux disease, obesity, diabetes and insomnia were seen to take a great part in connection to the migraine condition. In order to assess how the correlation between the disorders was relevant, p-values were computed to see which pairs of conditions were the most relevant within the different divisions of patients. The results for women and men presented some noticeable changes, from the frequency at which combinations of comorbid conditions appeared, to the number of disorders that presented a higher correlation among each other. This could be seen mainly through the number of disorders that had a significant p-value within the heatmaps associated with dendrograms: for the migraine female patients, the number of associations among comorbid conditions that were deemed as relevant within this set of patients was higher than what was seen through male migraine patients. Meaning that these relations among the comorbid conditions were more relevant for this group of patients than for their male counterparts. The analysis of networks related to women also offered a more dispersed and higher number of conditions, meaning that there is a higher number of conditions affecting this migraine population, in relation to men.

Using specific features of migraine patients such as the most common comorbidities, demographics of these patients such as age and gender, and even complexity measures such as the total number of comorbidities and ICD codes associated with each patient, as well as the number of admissions, it was possible to distinguish 4 different clusters. By performing Hierarchical Clustering methods, some differences were found among migraine patients. As to confirm and deepen the study of which features were of most importance in order to divide the migraine population, a principal component analysis was performed. For the totality of the principal components, the first three components explained 72% of the variance of the features. The total number of ICD codes, number of admissions, and number of comorbid diseases ranked highest in the assessment of the principal component, meaning that these explain the big majority of the variance among this principal component. Aspects found in the literature were seen through the results of the clusters, namely, Cluster 2, which comprised a percentage of almost 90% of female patients, with a median age of 31 years old. This cluster has female patients that fall into the childbearing ages and follows what is seen in the age-related prevalence of women having the highest percentage along the fertile period of time of ages from adulthood to menopause. It also can be traced back to migraine being related to women's sex hormones as this period of time in where women have a higher oscillation of these hormones due to menstruation, seen through Vetvik and MacGregor 2017 and Sacco et al. 2012.

A simpler approach for the eICU-CRD dataset was performed, while the analysis has brought up some changes from the MIMIC-IV dataset results. This can be explained through the context in which patients from eICU-CRD are inserted into, being that critical care units are places in which patients are in extreme conditions and thus may not translate to the true reality of migraine patients. The amount of migraine comorbid disorders was reduced in migraine patients, and the associations were found to follow what was seen in MIMIC-IV. Anxiety disorder, hypothyroidism, diabetes, stroke and obesity were major comorbid conditions found among these patients. The groups found among these patients were four. Interesting findings in one of the groups related to 90% of female patients, who had a high prevalence of the stroke condition and were in their childbearing ages. This links to what was seen through literature, similar to what happened in MIMIC-IV.

5.2 Limitations and future work

For the majority of phenotypic diseases networks found in literature, such as Kalgotra et al. 2017 and Hidalgo et al. 2009, in order to identify the patients' conditions, the ICD codes were used and simplified the process. However, adding additional information in order to have a more accurate representation of patients' conditions could have been used, resorting to notes of the datasets, to confirm the diagnosis. This can be a limitation to the phenotyping process, as ICD system may be used for billing purposes and adding multiple sources to understand the conditions of individuals can be of benefit and suggested by Shivade et al. 2014. Basing the diagnosis of patients using multiple sources eliminates the potential erroneous and imprecise information about patients.

Another limitation of this study is the fact that only one of the clustering methods was performed. Since Agglomerative Hierarchical Clustering has been widely used, while using Gower's distance for mixed type of data, this method was the one that was performed. However, it could be of interest to understand how other groups can be formed while using other clustering methods that have not been explored yet for migraine patients.

It could have been useful to incorporate time-series into this analysis, as both datasets contain enough information to understand the timeline of patients' diagnoses and see the temporal evolution of these diagnoses similarly to what has been done by Hidalgo et al. 2009. This could show some interesting age-related factors that have not been explored and could be key to understand migraine patients, due to the differences in population prevalence and its dependence on age.

Gender differences found among migraine patients should be further analyzed. There are some societal differences for both women and men and these should be taken into account, as the real prevalence of men with this condition may also be distorted, as there is a gender data gap for men's mental health, as seen through Al-Hassany et al. 2020. The ratio of women and men who report experiencing the migraine condition can be linked to the under-reporting of men, since this condition has been seen as a ”feminine” one and this could be reflected through epidemiological studies.

For the majority of the explored analysis, the key aspects have been the gender and age demographics of patients, and their comorbid conditions. However, other biological data could have been used, such as ethnicity, lifestyle-related features such as smoking, drug of alcohol habits. Information about medication could have the possibility to be incorporated and give insight to patients. It was thought to incorporate ethnicity as a feature in order to understand any separation among these patients. However, the results of the clustering and the SS when incorporating this feature were considered to be poor and could lead to a not real interpretation of reality. This originated from the fact that ethnicity is not well represented in the group of patients within the datasets, being that most of the patients in the data sets are white. Thus, as future work, incorporating ethnicity for this study could be an interesting component, there are some differences worth investigating.

Another important limitation of this study is the reducing gender concept. Only female and male patients were selected and identified. However, in the health research, it could also benefit from the inclusion of all individuals, giving access to a wider population while promoting gender equality (Williams et al. 2021).

Apart from that, the analysis was performed to hospitals in America, so it could be interesting to study other datasets from other countries, as different results could be found.

Bibliography

A. Ahmad and S. S. Khan. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7: 31883–31902, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2903568. URL http://arxiv.org/abs/1811.04364. arXiv:1811.04364 [cs, stat].

L. Al-Hassany, J. Haas, M. Piccininni, T. Kurth, A. Maassen Van Den Brink, and J. L. Rohmann. Giving Researchers a Headache - Sex and Gender Differences in Migraine. *Frontiers in Neurology*, 11: 549038, 2020. ISSN 1664-2295. doi: 10.3389/fneur.2020.549038.

C. Altamura, I. Corbelli, M. de Tommaso, C. Di Lorenzo, G. Di Lorenzo, A. Di Renzo, M. Filippi, T. B. Jannini, R. Messina, P. Parisi, V. Parisi, F. Pierelli, I. Rainero, U. Raucci, E. Rubino, P. Sarchielli, L. Li, F. Vernieri, C. Vollono, and G. Coppola. Pathophysiological Bases of Comorbidity in Migraine. *Frontiers in Human Neuroscience*, 15:640574, Apr. 2021. ISSN 1662-5161. doi: 10.3389/fnhum.2021.640574. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8093831/.

P. Amiri, S. Kazeminasab, S. A. Nejadghaderi, R. Mohammadinasab, H. Pourfathi, M. Araj-Khodaei, M. J. M. Sullman, A.-A. Kolahi, and S. Safiri. Migraine: A Review on Its History, Global Epidemiology, Risk Factors, and Comorbidities. *Frontiers in Neurology*, 12, 2022. ISSN 1664-2295. doi: 10.3389/fneur.2021.800605. URL https://www.frontiersin.org/article/10.3389/fneur.2021.800605.

A. M. Association. The differences between icd 9 and icd 10., 2015. URL www.ama-assn.org.

J. M. Banda, M. Seneviratne, T. Hernandez-Boussard, and N. H. Shah. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annual Review of Biomedical Data Science*, 1(1):53–68, July 2018. ISSN 2574-3414, 2574-3414. doi: 10.1146/annurev-biodatasci-080917-013315. URL https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-080917-013315.

M. Berg, Y. Appelman, M. Bekker, M. Blüm, A. Bos, J. Crasborn, B. Fauser, T. Goldhoorn, I. Hattem, I. van der Horst-Bruinsma, I. Klinge, E. Laan, L. Leliveld, A. Lagro-Janssen, S. Lo Fo Wong, A. Maas, A. Brink, J. Van Mens-Verhulst, A. Merens, and A. Toppen. *Gender and Health Knowledge Agenda*. May 2015. doi: 10.13140/RG.2.1.4359.0880.

J. Bezdek and N. Pal. Some new indexes of cluster validity. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 28:301–15, Feb. 1998. doi: 10.1109/3477.678624.

S. Bruehl, K. R. Lofland, E. M. Semenchuk, L. A. Rokicki, and D. B. Penzien. Use of Cluster Analysis to Validate IHS Diagnostic Criteria for Migraine and Tension-Type Headache. *Headache: The Journal of Head and Face Pain*, 39(3):181–189, Mar. 1999. ISSN 0017-8748, 1526-4610. doi: 10.1046/j.1526-4610.1999.3903181.x. URL http://doi.wiley.com/10.1046/j.1526-4610.1999.3903181.x.

Y. H. Chen, M. Karimi, and M. P. M. H. Rutten-van Mölken. The disease burden of multimorbidity and its interaction with educational level. *PloS one*, 15(12):e0243275, 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0243275.

A. Chmiel, P. Klimek, and S. Thurner. Spreading of diseases through comorbidity networks across life and gender. *New Journal of Physics*, 16(11):115013, Nov. 2014. ISSN 1367-2630. doi: 10.1088/1367-2630/16/11/115013. URL https://iopscience.iop.org/article/10.1088/1367-2630/16/11/115013.

S. Chu, Z. Wu, Z. Wu, J. Wu, and Y. Qian. Association Between Insomnia and Migraine Risk: A Case–Control and Bidirectional Mendelian Randomization Study. *Pharmacogenomics and Personalized Medicine*, Volume 14:971–976, Aug. 2021. ISSN 1178-7066. doi: 10.2147/PGPM.S305780.

M. Coscia. The Atlas for the Aspiring Network Scientist. Technical Report arXiv:2101.00863, arXiv, Feb. 2021. URL http://arxiv.org/abs/2101.00863. arXiv:2101.00863 [physics] type: article.

F. M. Cutrer, Z. Bajwa, and A. Sabahat. Pathophysiology, clinical manifestations, and diagnosis of migraine in adults. *Up To Date.[Online]*, 2012.

S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):54, Dec. 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0217-0. URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0217-0.

P. Diehr, G. Diehr, T. Koepsell, R. Wood, K. Beach, B. Wolcott, and R. K. Tompkins. Cluster analysis to determine headache types. *Journal of Chronic Diseases*, 35(8):623–633, Jan. 1982. ISSN 00219681. doi: 10.1016/0021-9681(82)90014-5. URL https://linkinghub.elsevier.com/retrieve/pii/0021968182900145.

A. R. Feinstein. The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of Chronic Diseases*, 23(7):455–468, Dec. 1970. ISSN 00219681. doi: 10.1016/0021-9681(70)90054-8. URL https://linkinghub.elsevier.com/retrieve/pii/0021968170900548.

J. C. Gower. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, 27(4):857–871, 1971. ISSN 0006-341X. doi: 10.2307/2528823. URL https://www.jstor.org/stable/2528823. Publisher: [Wiley, International Biometric Society].

C. Harrison, M. Fortin, M. van den Akker, F. Mair, A. Calderon-Larranaga, F. Boland, E. Wallace, B. Jani, and S. Smith. Comorbidity versus multimorbidity: Why it matters. *Journal of Comorbidity*, 11:2633556521993993, Mar. 2021. ISSN 2235-042X. doi: 10.1177/2633556521993993. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7930649/.

K. Hayrinen, K. Saranto, and P. Nykanen. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International Journal of Medical Informatics*, 77(5):291–304, May 2008. ISSN 13865056. doi: 10.1016/j.ijmedinf.2007.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1386505607001682.

C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Computational Biology*, 5(4):e1000353, Apr. 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000353. URL https://dx.plos.org/10.1371/journal.pcbi.1000353.

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. Mimic-iv, 2021. URL https://physionet.org/content/mimiciv/1.0/.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, Apr. 2016. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2015.0202. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202.

I. Jones, F. Cocker, M. Jose, M. Charleston, and A. L. Neil. Methods of analysing patterns of multimorbidity using network analysis: a scoping review. *Journal of Public Health*, Jan. 2022. ISSN 1613-2238. doi: 10.1007/s10389-021-01685-w. URL https://doi.org/10.1007/s10389-021-01685-w.

P. Kalgotra, R. Sharda, and J. M. Croff. Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *International Journal of Medical Informatics*, 108:22–28, Dec. 2017. ISSN 13865056. doi: 10.1016/j.ijmedinf.2017.09.014. URL https://linkinghub.elsevier.com/retrieve/pii/S138650561730237X.

J. H. Kim, K. Y. Son, D. W. Shin, S. H. Kim, J. W. Yun, J. H. Shin, M. S. Kang, E. H. Chung, K. H. Yoo, and J. M. Yun. Network analysis of human diseases using Korean nationwide claims data. *Journal of Biomedical Informatics*, 61:276–282, June 2016. ISSN 15320464. doi: 10.1016/j.jbi.2016.05.002. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046416300326.

T. Kurth, H. Chabriat, and M.-G. Bousser. Migraine and stroke: a complex association with clinical implications. *The Lancet. Neurology*, 11(1):92–100, Jan. 2012. ISSN 1474-4465. doi: 10.1016/S1474-4422(11)70266-6.

M. Leonardi and A. Raggi. A narrative review on the burden of migraine: when the burden is the impact on people's life. *The Journal of Headache and Pain*, 20(1):41, Dec. 2019. ISSN 1129-2369, 1129-2377. doi: 10.1186/s10194-019-0993-0. URL https://thejournalofheadacheandpain.biomedcentral.com/articles/10.1186/s10194-019-0993-0.

N. Menachemi and T. H. Collum. Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, 4:47–55, May 2011. ISSN 1179-1594. doi: 10.2147/RMHP.S12985. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3270933/.

J. M. Pavlovic, D. Akcali, H. Bolay, C. Bernstein, and N. Maleki. Sex-related influences in migraine: Sex-Related Influences in Migraine. *Journal of Neuroscience Research*, 95(1-2):587–593, Jan. 2017. ISSN 03604012. doi: 10.1002/jnr.23903. URL `https://onlinelibrary.wiley.com/doi/10.1002/jnr.23903`.

O. J. Pellicer-Valero, C. Fernández-de-las Peñas, J. D. Martín-Guerrero, E. Navarro-Pardo, M. I. Cigarán-Méndez, and L. L. Florencio. Patient Profiling Based on Spectral Clustering for an Enhanced Classification of Patients with Tension-Type Headache. *Applied Sciences*, 10(24):9109, Dec. 2020. ISSN 2076-3417. doi: 10.3390/app10249109. URL `https://www.mdpi.com/2076-3417/10/24/9109`.

S. A. Pendergrass and D. C. Crawford. Using Electronic Health Records to Generate Phenotypes for Research. *Current protocols in human genetics*, 100(1):e80, Jan. 2019. ISSN 1934-8266. doi: 10.1002/cphg.80. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6318047/`.

B. L. Peterlin, S. Gupta, T. N. Ward, and A. MacGregor. Sex Matters: Evaluating Sex and Gender in Migraine and Headache Research. *Headache*, 51(6):839–842, June 2011. ISSN 0017-8748. doi: 10.1111/j.1526-4610.2011.01900.x. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3975603/`.

T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, 5:180178, Sept. 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.178.

N. Reddy, M. N. Desai, A. Schoenbrunner, S. Schneeberger, and J. E. Janis. The complex relationship between estrogen and migraines: a scoping review. *Systematic Reviews*, 10(1):72, Dec. 2021. ISSN 2046-4053. doi: 10.1186/s13643-021-01618-4. URL `https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-021-01618-4`.

V. Regitz-Zagrosek. Sex and gender differences in health: Science & Society Series on Sex and Science. *EMBO reports*, 13(7):596–603, July 2012. ISSN 1469-221X, 1469-3178. doi: 10.1038/embor.2012.87. URL `https://onlinelibrary.wiley.com/doi/10.1038/embor.2012.87`.

S. Sacco, S. Ricci, D. Degan, and A. Carolei. Migraine in women: the role of hormones and their impact on vascular diseases. *The Journal of Headache and Pain*, 13(3):177–189, Apr. 2012. ISSN 1129-2369, 1129-2377. doi: 10.1007/s10194-012-0424-y. URL `https://thejournalofheadacheandpain.biomedcentral.com/articles/10.1007/s10194-012-0424-y`.

M. Schurks, P. M. Rist, M. E. Bigal, J. E. Buring, R. B. Lipton, and T. Kurth. Migraine and cardiovascular disease: systematic review and meta-analysis. *BMJ*, 339(oct27 1):b3914–b3914, Oct. 2009. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.b3914. URL `https://www.bmj.com/lookup/doi/10.1136/bmj.b3914`.

M. Schürks, J. E. Buring, and T. Kurth. Migraine features, associated symptoms and triggers: A principal component analysis in the Women's Health Study. *Cephalalgia*, 31(7):861–869, May 2011. ISSN

0333-1024, 1468-2982. doi: 10.1177/0333102411401635. URL http://journals.sagepub.com/doi/10.1177/0333102411401635.

C. Shivade, P. Raghavan, E. Fosler-Lussier, P. J. Embi, N. Elhadad, S. B. Johnson, and A. M. Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, Mar. 2014. ISSN 1067-5027, 1527-974X. doi: 10.1136/amiajnl-2013-001935. URL https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-001935.

S. E. Short, Y. C. Yang, and T. M. Jenkins. Sex, Gender, Genetics, and Health. *American Journal of Public Health*, 103(Suppl 1):S93–S101, Oct. 2013. ISSN 0090-0036. doi: 10.2105/AJPH.2013.301229. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786754/.

Steiner, T. J. Stovner, L. J. Jensen, R. Uluduz, and D. Katsarava. Migraine remains second among the world's causes of disability, and first among young women: findings from GBD2019. *The Journal of Headache and Pain*, 21(1):137, s10194–020–01208–0, Dec. 2020. ISSN 1129-2369, 1129-2377. doi: 10.1186/s10194-020-01208-0. URL https://thejournalofheadacheandpain.biomedcentral.com/articles/10.1186/s10194-020-01208-0.

J. P. Sturmberg, L. O. Getz, K. C. Stange, R. E. Upshur, and S. W. Mercer. Beyond multimorbidity: What can we learn from complexity science? *Journal of Evaluation in Clinical Practice*, 27(5):1187–1193, 2021. ISSN 1365-2753. doi: 10.1111/jep.13521. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jep.13521. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jep.13521.

G. E. Tietjen, N. A. Herial, J. Hardgrove, C. Utley, and L. White. Migraine Comorbidity Constellations. *Headache: The Journal of Head and Face Pain*, 47(6):857–865, June 2007. ISSN 00178748, 15264610. doi: 10.1111/j.1526-4610.2007.00814.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1526-4610.2007.00814.x.

M. C. Tonini. Gender differences in migraine. *Neurological Sciences*, 39(S1):77–78, June 2018. ISSN 1590-1874, 1590-3478. doi: 10.1007/s10072-018-3378-2. URL http://link.springer.com/10.1007/s10072-018-3378-2.

J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland. Defining Comorbidity: Implications for Understanding Health and Health Services. *The Annals of Family Medicine*, 7(4):357–363, July 2009. ISSN 1544-1709, 1544-1717. doi: 10.1370/afm.983. URL http://www.annfammed.org/cgi/doi/10.1370/afm.983.

K. G. Vetvik and E. A. MacGregor. Sex differences in the epidemiology, clinical features, and pathophysiology of migraine. *The Lancet. Neurology*, 16(1):76–87, Jan. 2017. ISSN 1474-4465. doi: 10.1016/S1474-4422(16)30293-9.

T. Vos, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, R. S. Abdulkader, and Abdulle. Global, regional, and national incidence, prevalence, and years lived with disability for

328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100):1211–1259, Sept. 2017. ISSN 01406736. doi: 10.1016/S0140-6736(17)32154-2. URL `https://linkinghub.elsevier.com/retrieve/pii/S0140673617321542`.

WHO. International classification of diseases (icd), 2022. URL `http://wwwho.int/classifications/icd/en/`.

A. Williams, J. S. Lyeo, S. Geffros, and A. Mouriopoulos. The integration of sex and gender considerations in health policymaking: a scoping review. *International Journal for Equity in Health*, 20(1):69, Dec. 2021. ISSN 1475-9276. doi: 10.1186/s12939-021-01411-8. URL `https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-021-01411-8`.

Y. W. Woldeamanuel, B. M. Sanjanwala, A. M. Peretz, and R. P. Cowan. Exploring Natural Clusters of Chronic Migraine Phenotypes: A Cross-Sectional Clinical Study. *Scientific Reports*, 10(1):2804, Feb. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-59738-1. URL `https://doi.org/10.1038/s41598-020-59738-1`.

World Health Organization. Medication safety in polypharmacy: technical report. Technical report, World Health Organization, Geneva, 2019. URL `https://apps.who.int/iris/handle/10665/325454`. Section: 61 p. WHO/UHC/SDS/2019.11.

R. Xu and D. C. Wunsch. Clustering Algorithms in Biomedical Research: A Review. *IEEE Reviews in Biomedical Engineering*, 3:120–154, 2010. ISSN 1937-3333, 1941-1189. doi: 10.1109/RBME.2010.2083647. URL `http://ieeexplore.ieee.org/document/5594620/`.