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

Maria Manuel Lopes Jacinto
maria.jacinto@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2022

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 directly related to patients with multimorbidity.

Keywords: Migraine, comorbid conditions, Electronic Health Records phenotyping, Network Science, Clustering.

1. Introduction

Migraine is a condition that affects the global population and is considered as one of the top ten most disabling conditions (Vos et al. 2017). This challenging condition carries a burden to healthcare and can lead to a poor life quality for the individuals affected by it (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 (Peterlin et al. 2011). Women's pre and post menopausal stages of life have different prevalences of this disorder, when compared to men. The gender differences associated to this condition were the motivation to deepen the study of migraine and characterize its patients (Berg et al. 2015).

Migraine does not often appear on its own (Altamura et al. 2021). 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 studied through Phenotypid Disease Networks (PDNs) (Hidalgo et al. 2009). This allows to unveil not so obvious links between conditions, as seen through 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 and some clustering approaches for comorbid conditions among these patients has been seen through Pellicer-Valero et al. 2020.

The goal of this dissertation is to select patients that experience the condition of Migraine and their most common comorbidities within the available datasets using Electronic Health Records (EHRs) 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 characteristics, for the sake of gathering essential information about them, as to be able to have a more personalized observation about this type of patients.

The datasets used was from Medical Information Mart for Intensive Care IV and eICU Collaborative Research Database (MIMIC-IV) (Johnson et al. 2021) and eICU Collaborative Research Database (eICU-CRD) (Pollard et al. 2018), which hold information about patients in the United States of America. The eICU-CRD contains data about patients in critical care units, while the MIMIC-IV dataset comprises hospitals' wide EHRs about all patients, including critical care units.

2. Background

The migraine condition shows several factors that can be differentiated for both women and men. 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 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. 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. 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.

EHRs 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 finding cohorts of patients that are associated to a certain disease or a desired characteristic and exploring within said population. These can be used coupled to different methods. Using a network science approach, in order to understand connections between disorders and trying to grasp more information from more complex concepts that are often too difficult to keep track. The concept of PDNs was firstly introduced by Hidalgo et al. 2009 and it is a tool that aids in the understanding of phenotype differences inside a population and gives insight about disease progression. In fact, PDNs 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. In this type of networks, the nodes usually represent the diseases, while the edges are the connections between said diseases.

Clustering has also been a method used in other 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 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.

One of the most important steps in clustering is the evaluation or validation of this method. 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. The equation can be found in 1, 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.

$$\text{Silhouette Score } (i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{1}$$

Clustering mixed type data is a challenging concept. As seen through Ahmad and Khan 2019 review, some studies have been done by applying Gower's distance while using agglomerative hierarchical clustering methods, in order to tackle these difficulties. Principal Component Analysis (PCA) is a method used in order to 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. 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.

3. Dataset Preliminary Analysis

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. 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. 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 lead in order to identify them.

One of the main differences between the presented datasets lays 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 EHRs, which can include information from laboratories and data from another type of specific hospital unit such as emergency departments. 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.

3.1. 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 EHRs, critical care units and emergency departments. This dataset is an updated version of an already released dataset called MIMIC-III, adding more information about patients in a structured and easier approach, while dividing clearly the origin of the collection of data. The collection of the medical data was followed through 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.

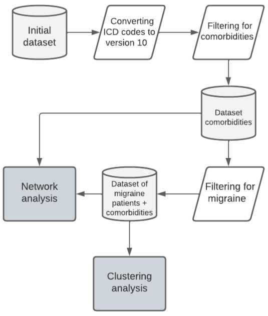

Initial dataset
Converting ICD codes to version 10
Filtering for comorbidities
Dataset comorbidities
Network analysis
Dataset of migraine patients + comorbidities
Filtering for migraine
Clustering analysis

Figure 1: Pipeline to obtain the prepared datasets for both network and clustering analyses.

3.2. eICU

The eICU-CRD dataset 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. 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.

3.3. Pre-processing the Datasets

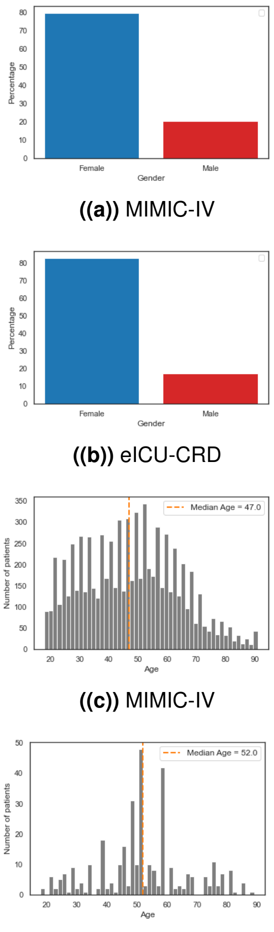
Prior to deepening the study and performing any type of analysis to both datasets, it was necessary to go through a pre-processing stage. In Figure 1, it is possible to assess which were the relevant steps in order to obtain all of these analysis. It was necessary to have the diagnoses codes to only contain one of the revisions. 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 to the most common comorbidities of migraine and migraine itself. By narrowing down the population to only patients who expressed one of the conditions, the number of patients is reduced and it becomes possible to study in more detail how these comorbidities are related to each other.

4. Results & discussion

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 displays a set of characteristics that are intrinsic to the condition. 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 for the MIMIC-IV dataset. Both genders in eICU-CDR have a similar



((a)) MIMIC-IV

((b)) eICU-CRD

((c)) MIMIC-IV

((d)) eICU-CRD

Figure 2: 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). Age histogram of migraine population in MIMIC-IV dataset in (c) and eICU-CRD in (d). It is possible to assess the median age in orange in each histogram.

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 2(c) and 2(d) respectively.

Since this dissertation focus on the most common comorbidities of migraine patients, it is of importance to understand the prevalence of said conditions inside this group. This can be seen in Figure 3 for the MIMIC-IV dataset, where the percentage of each of the most common comorbidities can be evalutated. 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%).

4.2. Correlation between Migraine comorbid conditions in MIMIC-IV patients

An analysis on how the most commonly occurring comorbidities within this group of patients co-occur was performed. This allows to study 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. Computing the sum of how many times each combination of two comorbidities appear 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.

4.3. 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 and can be see in Figure 4. The nodes represent the comorbid conditions and the interactions between each other, meaning how many times they co-occur together, can be seen through the edges. The wider the edge, the more times the combination of these two conditions occurs, meaning that the stronger 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 above 1% of the population. Hereby, the combinations of the two comorbidities had to account

[1] `https://networkx.org/`

[2] `https://gephi.org/`

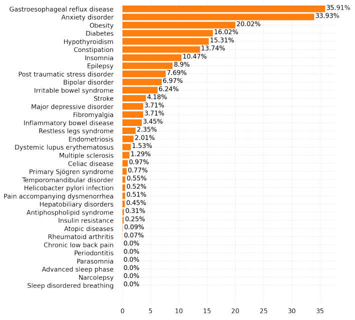


Figure 3: Comorbidities related to migraine and their percentage in the subgroup of patients that are affected by this condition.

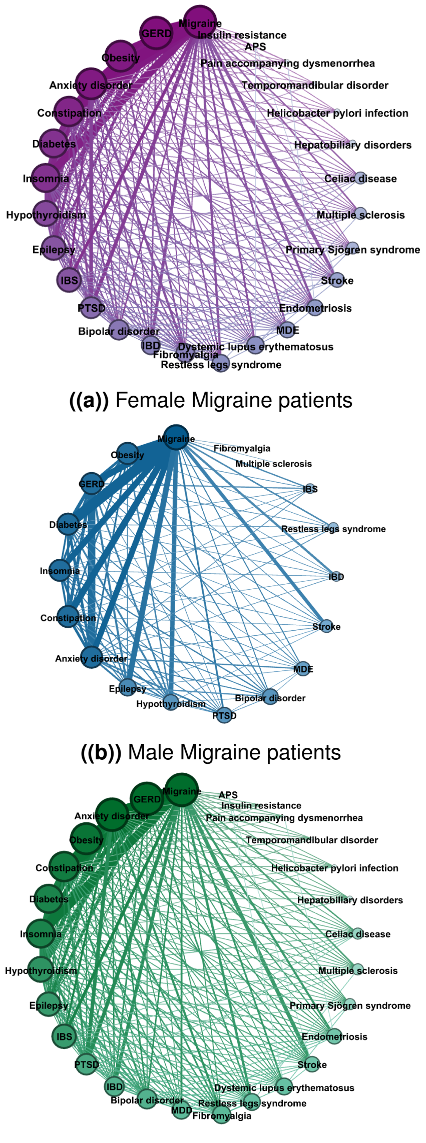
for at least 1% of the combinations in these graphs. The networks that are presented have been adjusted to display a circular layout, and Gephi's software allows to compute the degree of these nodes, making it possible to order them through this measurement and in a counter clockwise direction.

For the MIMIC-IV population, it was possible to verify through Figure 4 (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. Regarding the differences between the female and male population, it is demonstrated through Figure 4 (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 (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 (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 related to female reproductive organ, and the menstrual cycle. Endometriosis and pain accompanying dysmenorrhea can be seen through women's graph, but not in men's.

4.4. 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 to 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 5 (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 of these comorbid conditions.

Dividing the MIMIC-IV population into female and male patients, allows 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 analysed through Figure 5 (c) and (d) respectively. These two heatmaps showed a similar outcome to the total population ones, seen in Figure 5 (a) and (b). As for the male population, it can be verified through Figure 5 , 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 male migraine population in (e), there is a noticeable change from the counterpart of female migraine population. All the associations between the comorbidities have a smaller count

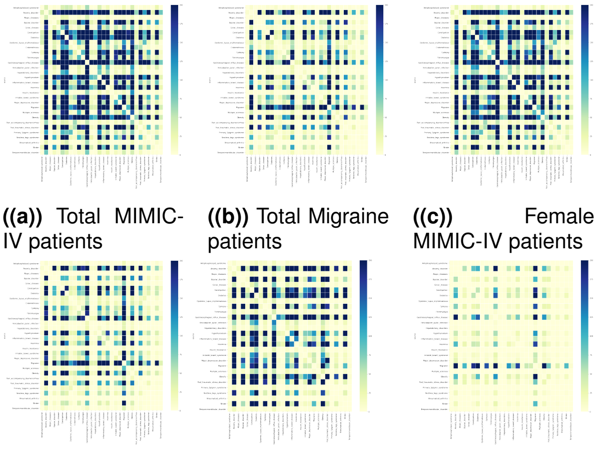


((a)) Female Migraine patients

((b)) Male Migraine patients

((c)) Total Migraine patients

Figure 4: Circular graphs representing how the most common comorbidities of migraine are related to each other in MIMIC-IV's migraine patients. GERD - Gastroesophageal reflux disease; IBD - Inflammatory bowel disease; IBS - Irritable bowel syndrome; MDE - Major depressive episode; APS - Antiphospholipid syndrome; PTSD - Post traumatic stress disorder.



((a)) Total MIMIC-IV patients **((b))** Total Migraine patients **((c))** Female MIMIC-IV patients

((d)) Female Migraine patients **((e))** Male MIMIC-IV patients **((f))** Male Migraine patients

Figure 5: Heatmap displaying the counts of two of the comorbid conditions associated to migraine, in this case for the whole **female** MIMIC-IV population and the **female** migraine subgroup. Darker shade of blue codes for a higher appearance, opposed to lighter shades of blue/white, coding for lowest values.

among men, and the values are lower overall.

4.5. Statistical relevance of comorbid associations

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 for each of the groups. The p-value of each combination of the two comorbid conditions was obtained resorting to the hypergeometric function of the *SciPy* Python's package. [3] It was pertinent to build networks while narrowing to pairs of associations that are associated to a p-value that is lower than 0.05. This allows for the understanding of which conditions are more associated to each other in each of these groups and subgroups that have been formed.

4.6. Network visualization of relevant comorbid associations

The networks 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. Regarding the network computed for the total population of migraine when comparing to the MIMIC-IV population who are associated to the comorbid conditions of migraine, we can verify in Figure 6, that for the most part, there are three groups in which the comorbidities are related to each other.

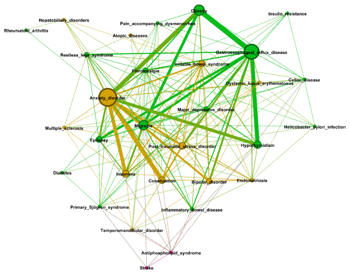


Figure 6: Representation of the graphs containing the most relevant combinations of comorbididities within the **total** population of the MIMIC-IV dataset and the **total** migraine population, which stated a p-value < 0.05.

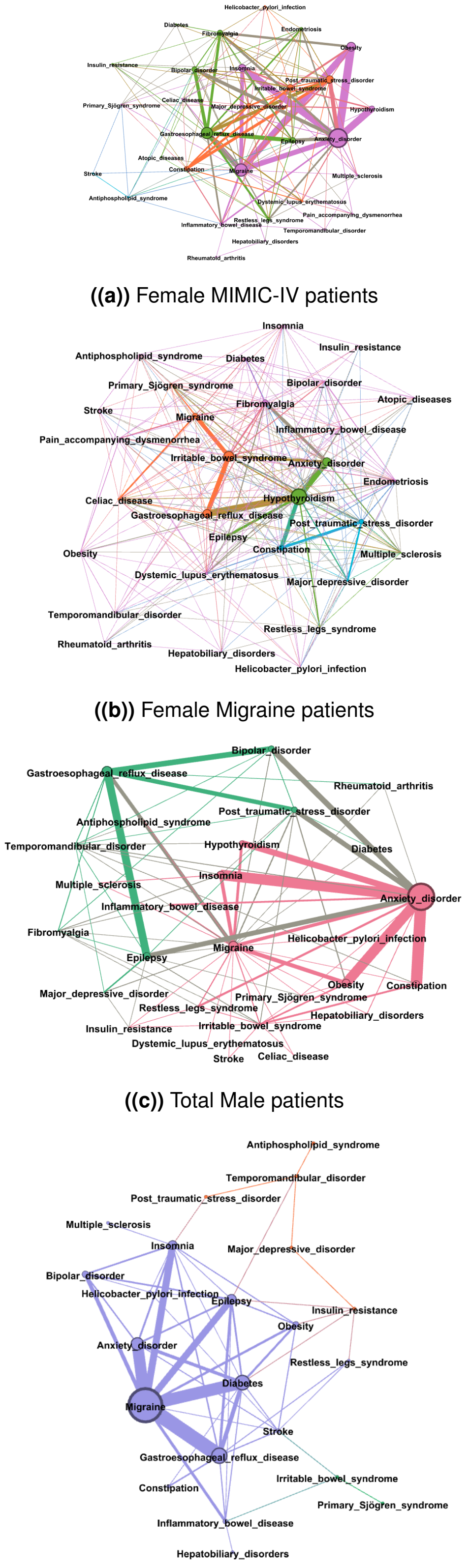
Regarding the networks that contain the female population, it is possible to verify a big difference in relation to the whole female population of MIMIC-IV and the migraine female subgroup. In Figure 7 (a), it is possible to see that some of the common comorbidities of migraine are often coupled together, 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 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. Comparing it with the subgroup of migraine, it is possible to identify much broader connections in subgroups of Figure 7 (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.

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. As for the totality of male patients in the MIMIC-IV dataset, we can see that in (c) 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.

4.7. 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.

[3] https://scipy.org/



((a)) Female MIMIC-IV patients

((b)) Female Migraine patients

((c)) Total Male patients

((d)) Male Migraine patients

Figure 7: 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.

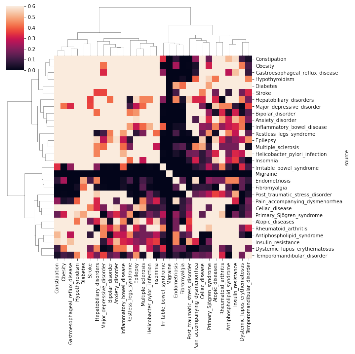


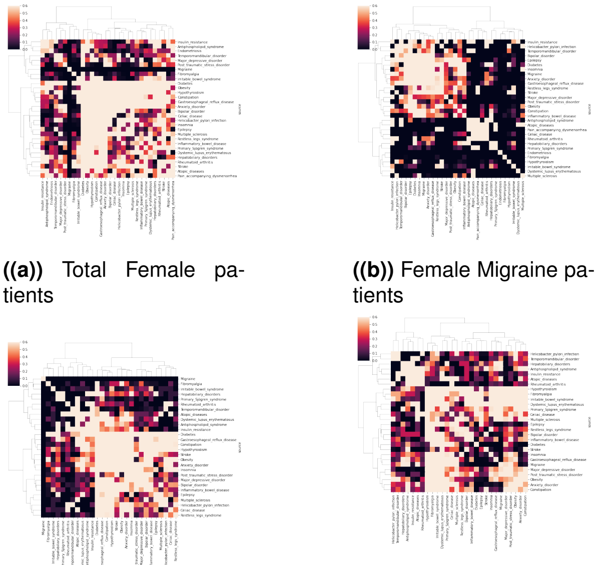
Figure 8: Dendogram with a heatmap of the **total** migraine population and displaying the probability of a patient having two of the comorbid conditions associated to migraine. A darker colour codes for a lower p-value, thus a higher statistical relevance of the association between two conditions.

The results depend, of course, of 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.

As seen in Figure 8, 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.

When it comes to the gender basis analysis, it was possible to have two different outcomes regarding each gender. When analysing the dendograms with a heatmaps related to women and their comorbid associations, it is possible to see in Figure 9 (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. As for Figure 9 (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.

As for the male population, as seen in Figure 9 (c) and (d), the most prominent correlations to migraine are different from what is seen in the female population's dendograms with heatmaps. Looking into Figure 9 (d), 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.



((a)) Total Female patients

((b)) Female Migraine patients

((c)) Total Male patients

((d)) Male Migraine patients

Figure 9: 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.

5. Identification of subgroups within migraine patients in MIMIC-IV

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. 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.

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, and a PCA.

5.1. 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. 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. In Table 1 demonstrates which are the characteristics of the population of each the obtained clusters for the MIMIC-IV population.

Looking at the most important differentiation that are age and the ratio of female to male patients, cluster 2 can be related to the gender differences found in migraine patients. During women's childbearing ages, women tend to have a higher prevalence of this disorder. This falls into what is seen in this cluster, meaning that it comprises mostly female patients in their most fertile ages and can 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. 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.

Table 1: 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

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.

5.2. Principal Components Analysis and Hierarchical Clustering

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. One important aspect that should be analysed 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 10, as it is possible to verify which of the features has contributed the most for each of the principal components. 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 ICD codes per patient, and the number of admissions pay 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. 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.

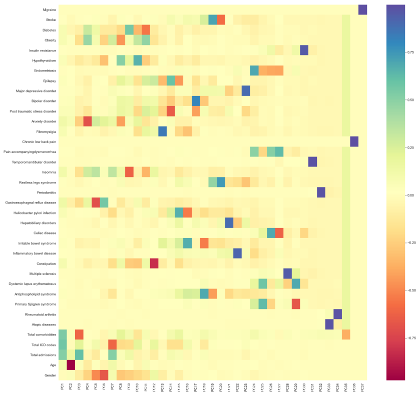


Figure 10: Generated correlation matrix plot between the features used for clustering and the principal components. Higher values are coded as blue, neutral values are yellow and lower values are represented in red.

6. 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.

Similarly 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. 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 analyse. Since the extreme conditions of patients in critical care units are difficult to manage, and often do not allow for a deeper understading 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 syndrom, helicobacter pylori infection, hepatobiliary disorders, celiac disease, irritable bowel syndrome, inflammatory bowel disease, multiple sclerosis, antiphospholipid syndrom, 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. The analysis perfomed for this dataset was simpler, and followed the same results obtained by the MIMIC-IV, with an emphasis in the relationship between stroke and migraine for women.

7. Conclusions

The presented work aims to characterize patients who are associated to the migraine condition and understand how their most common comorbid disorders are associated to each other. It was possible to understand how the most common comorbidities associated to 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 to it, and several already known gender differences among patients.

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. The analysis of the networks related to women 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 to 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. 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 literature were seen through the results of the clusters.

A simpler approach for the eICU-CRD dataset was performed, while the analysis has brought up some significant 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 the true reality of migraine patients.

As a future work hypothesis, adding additional information in order to have a more accurate representation of patients' conditions could have been used, resorting to notes of the datasets or even medication. This can be a limitation to the 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 suggest by Shivade et al. 2014. Studying other non-American datasets could also be of interest when performing these analyses.

References

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. arXiv:1811.04364 [cs, stat].

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.

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.

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.

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.

A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark. Mimic-iv, 2021.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.