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

Maria Luísa Marote e Costa, luisamarote@tecnico.ulisboa.pt[1]

[1]Instituto Superior Técnico, University of Lisbon, Portugal

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

***Index Terms*—Multimorbidity, Dementia, Markov Chains, Temporal Sequence Alignment, Clustering**

I. Introduction

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

As the care providing system improves, the population in general will tend to live longer. With this growth in ageing, the incidence of multimorbidity will tend to increase, leading to an overload of the healthcare system. Given the fact that these patients need regular attention, it raises a need to prepare treatment plans to serve their needs. The growing attention towards the process of data mining supports this search for insightful information. Data mining is focused on discovering patterns and correlations within heterogeneous large data sets, resorting to a broad range of techniques which can include machine learning and artificial intelligence.

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

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

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

II. Concepts and related Work

Electronic Health Records (EHRs) can be associated to the appropriate evidence-based tools, having a lot of potential

when it comes to developing clinical decision support systems for healthcare providers, which could have a huge impact on patient diagnosis, as well as in quality measures. Furthermore, EHRs can be used to assist in a broad range of experiments regarding patient information, such as discovering phenotype-genotype associations, establishing clinical trial protocols, automating drug event detection, as well as prevention, accelerate the research on precision medicine [2], and electronic phenotyping, which can be defined as finding patients with specific conditions or outcomes [3], representing one of the major approaches concerning the utilization of EHRs.

Ageing is the most persistent factor mentioned in the literature when it comes to the increase in multimorbidity risk, having several studies mention a direct association between age and prevalence of MM, pointing as well that, through their systematic review on studies concerning MM, most individuals older than 65 years live with MM [4]. However, it is important to advert to the fact that, despite of the increase of this prevalence along with ageing, it is not a condition only affecting the elder population [5]. The increase of MM amongst the population has also been associated to lifestyle changes, health seeking behavior and the environment [6].

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

One possible approach to finding these patterns is by analysing their clinical pathways [7]. By resorting to Markov Chain models, we can explore a probabilistic model that is able to identify patterns in clinical pathways on the population level, as was done in [8]. Markov Chain models consider patients in discrete states, and events represent transitions from one state to another [9]. We adapt this notion of discrete states to represent different medical speciality consults. For instance, [10] resorts to Markov Models in order to find patterns in a certain set of patients regarding chronic illnesses, focusing as well on predicting which illness a patient is more likely to be affected by in the future.

Another possible approach to address the heterogeneity amongst clinical populations and how to target it, is through patient stratification [11]. To this end, AliClu – an algorithm which combines temporal sequence alignment and hierarchical clustering – was developed and applied on a set of Rheumatology patients, in order to stratify them based on their medication switches throughout time [12]. More specifically, AliClu starts by using the Temporal Needleman-Wunsch (TNW) procedure to align sequences with temporal information between events. Then, a hierarchical clustering method is implemented, resorting to pairwise scores obtained during the alignment process [12]. This algorithm has the advantage of being easily adapted to analyse any kind of temporal events, as to identify clusters of patients that show similar transitions between medical consults, both in terms of discrete sequences of events and the time lag between the existing switches, allowing the identification of common patterns amongst a group of patients. Similarly to this set of methods used to tackle the stated concerns, [13] published a population based study that aims at identifying temporal patterns in patient disease trajectories. To that end, firstly, pairwise disease associations as well as their temporal directionality are assessed, moving on to a Dynamic Time Warping (DTW) based clustering algorithm. This DTW technique is applied on the common disease trajectories in order to group them according to shared temporal patterns.

According to the World Health Organization (WHO), Dementia is defined as a syndrome of chronic or progressive nature, leading to a degradation in cognitive function and affecting memory, the capacity to process thought, orientation, language, judgement, ability to calculate and learn, within others [14]. A phenotypic heterogeneity regarding patients with Dementia has been previously flagged, mainly due to the fact that this disease can result from a number of distinct factors and other conditions, with diverse etiology and pathophysiology [15]. Within patients with Dementia, various symptoms, disease trajectories, as well as individual onset ages can be identified [16]. Given all these challenges and the importance of better understanding these patients, we selected them as our study cohort.

Altogether, based on the stated observations concerning the lack of a strong patient-centered care giving, our aim is to focus on clinical pathway analysis, resorting to Markov chains and the AliClu algorithm, coupled with a phenotype screening, in order to more accurately characterize Dementia patients and identify distinguishable patterns amongst them.

III. Methodology

In this study, a pipeline was developed to identify characteristics and patterns within Dementia patients, regarding their phenotypes and clinical pathways. The pipeline includes the following steps: (i) initial phenotype screening to characterise the dataset; (ii) creation of transition matrices to identify most common medical consult activity; (iii) clustering algorithm based on medical consult pathway; (iv) characterisation, visualisation and phenotype screening of the obtained clusters, including age, gender, chronic diseases, medication, hospitalization and emergency analysis.

A. Available data and initial phenotyping

In order to fulfil our goals and motivation, data was collected from Hospital da Luz Lisboa, dated from January 2007 to August 2021. Collected data included $302,709$ patients, $63,786$ of them suffering from multimorbidity. Among these, $1,924$ (114 female and 777 male) were identified as having Dementia. Data concerning these patients' age, gender, chronic diseases, as well as information on $20,033$ medical consults attended, was gathered. An initial approach involved characterisation of the data set in terms of the collected phenotypes, having observed distributions for the whole population and by gender, in order to better understand and characterize the data set in hands. An analysis of the population age, amount and types of chronic diseases, as well as the medical specialities involved was carried out.

Regarding the clinical pathway analysis phase, it is important to point out the fact that within the 1924 patients with Dementia present in the study cohort, 59 of them did not have information on their medical consults. For this reason, it was not possible to include these patients in this part of the study, having remained 1865 Dementia patients for the clinical pathway analysis carried out. Furthermore, in order to accomplish the medication and hospital admissions analysis, data on $35,263$ prescriptions given to Dementia patients, $2,078$ hospital admissions (HA) and $9,991$ emergency episodes (EE) was made available for this purpose. Information relative to patients who did not integrate any of the obtained clusters was mapped out of the data sets and the remaining data was mapped to each of the clusters.

B. Markov chains

The following steps are mainly centered on hospital activity, in order to identify recurring patterns of medical consults within Dementia patients. Through estimation of Markov chains, it was possible to determine the most prevalent transitions between consults. This was accomplished by formulating a transition matrix (TM), which shows the transition probabilities between two states (i.e., consults). Given a square matrix of all possible medical specialities, we can calculate the conditional probabilities of moving to a second speciality consult (*j*), given the previous one (*i*). This is achieved by dividing the number of times that each transition occurs by the prevalence of the origin medical speciality consult, filling the transition matrix as follows:

$$TM_{ij} = P(i \rightarrow j) = \frac{\#(i \rightarrow j)}{\#i} \tag{1}$$

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

C. AliClu algorithm

In order to stratify Dementia patients considering their activity patterns within the hospital, the AliClu algorithm was used. However, slight adaptations considering the data in hands, which will be explained further ahead, were necessary, both on the algorithm itself, as well as in the pre processing step. In summary, first of all, the available data from HLL concerning consult activity of patients with Dementia was converted from panel data format to the appropriate sequences to be used as input for AliClu. Then, an optimization process was implemented in order to reach the best stratification possible and lastly, the final clusters were obtained.

1) Preprocessing: Converting the medical consult data from panel data format to the correct input temporal sequences, named prefix-encoded (PE) sequences, involved the following steps: (1) each event, in this case medical consult, is assigned a label, (2) the time elapsed between consecutive events is calculated and fixed in the sequence in days and (3) filter patients who do not fit the appropriate criteria to undergo the clustering algorithm. Specifically, in view of the fact that the focus is on patients with multimorbidity, patients who only have one medical consult present in their temporal sequence, removing these patients is an intuitive step. Furthermore, in order to avoid outliers as much as possible, an additional filtration is done, removing all patients who attend a number of different consults superior to a certain threshold. This threshold is given by the 95 percentile regarding the number of different consults attended throughout a patient's pathway.

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

2) Parameter optimization: AliClu was designed to return one set of clusters for the best combination of gap penalty (g), and number of clusters (k), being the temporal penalty (Tp), established in the beginning of the process and hence, not iterated during the development. This is a very sensitive algorithm when it comes to the choice of these parameters, meaning that a slight change will deliver completely different results, which underlines the importance of tuning these parameters in the most efficient way possible. Hence, the algorithm was adapted to, based on a clustering index called silhouette score (SS), which is a measure of how close an element is to it's own cluster, compared to how close it is to the other clusters, decide on the set of parameters that will deliver the best subgroups. So, AliClu was adapted to return a set of clusters for each combination of the three parameters. For each set of clusters, the average SS was computed, since this metric is a good indicator of the cluster's content. From the collection of average silhouette scores obtained for each set of clusters, the optimum parameters were chosen by searching for the highest score obtained.

For the purpose of searching through the broader range of parameters possible, considering the data set in hands, an initial implementation was carried out, serving the sole purpose of deciding on the best set of parameters. Hence, the whole dataset was split into six partitions of equal size and an adaptation of the original AliClu algorithm was applied to those partitions. This implementation returns, for each partition, the average SS obtained considering each set of parameters (*g*, *Tp*, *k*) analysed. These are the most crucial algorithm parameters to take into consideration, however, assumptions were made regarding other aspects that had to be verified, based once again in average SS results.

Firstly, it was hypothesized that the best linkage function used for the hierarchical clustering process of the algorithm was Ward's method. It was then confirmed that the single, complete and average methods were not indeed better. Second of all, it was verified if a number of bootstrap samples higher than 250 did not lead to better outcomes. Lastly, we assumed that gap penalties outside the range of -0.5 to 0.5 were not appropriate, hence, those gaps from -1 to -0.6 and 0.6 to 1 were tested to prove this hypothesis. It is important to point out that, to support this SS analysis, the obtained clusters in each optimization step were visually analysed so to guarantee that the changes in the SS results went accordingly to what

was observed.

3) Obtaining the final clusters: Taking the results from the parameter optimization process, which gave the values for the gap and temporal penalty that lead to the best data partition, these were set to run the algorithm for the whole Dementia data set. Regarding the number of clusters, this is a parameter that is sensitive to the amount of data, hence the optimum k was chosen in this phase.

D. Cluster analysis and phenotyping

Subsequently to finding the optimum set of parameters which lead to the best patient clusters, it is imperative to assess their quantitative and qualitative reliability. To this end, through every phase of the parameter optimization process, a manual observation of the so considered optimum groups obtained was done, in order to confirm that the algorithm was indeed grouping patients which presented a similar pathway regarding consult attendance. Moreover, besides the average silhouette scores calculated across all samples, the mean silhouette was also calculated within each cluster, so to assess the similarity of the elements within a certain cluster, when compared to elements allocated to different groups. Additionally, AliClu has a cluster stability analysis process integrated where, for each of the final clusters, three metrics of three different indices are calculated, which are the median, average and standard deviation of the Jaccard Index, Dice Coefficient and Recovery Rate. By analysing these metrics, it is possible to conclude about the quality of each developed cluster, considering that the average values should be as close to 1 as possible, while the standard deviations should be closer to 0.

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

E. Medication analysis

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

The next step was identifying the most prevailing medication within the cohort and each cluster. However, due to the vast amount of medicines identified in the dataset and knowing that several ones have the same purpose, we grouped the medication present in the cohort by class. The most relevant medication classes regarding the cohort under study were identified and the different medicines were assigned to them, having performed an analysis not of a medicine itself, but of each of the prevailing classes of medication, both for the whole dataset and for each individual cluster.

F. Hospital admission and emergency analysis

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

An OR represents a measure of association between an exposure and an outcome, meaning, it gives the odds of an outcome occurring given a certain exposure, when compared to the odds of occurring when that exposure does not exist [17]. A value of OR equal to one means that the exposure does not influence the outcome, a value lower than one indicates that the exposure causes lower odds of the outcome occurring, while an OR higher than one implies that the exposure leads to higher outcome odds. In our specific case, the goal is to understand if a patient belonging to a specific cluster (exposure) is more or less probable of being admitted or having an emergency episode (outcome).

The formula for calculating an OR is the following:

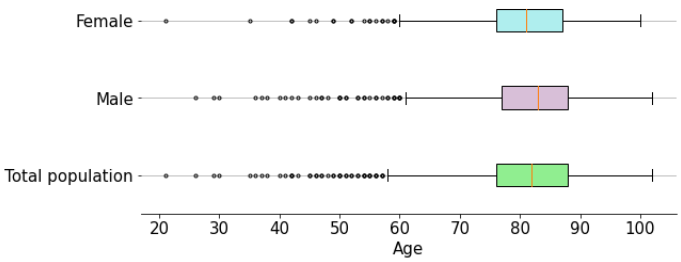
$$OR = \frac{a/c}{b/d} \tag{2}$$

where *a* represents the number of exposed cases, *b* the number of exposed non-cases, *c* the number of unexposed cases and *d* the number of unexposed non-cases. In this specific case, aiming at calculating the odds ratio for a patient in a certain cluster being admitted to the hospital or having an emergency episode, *a* corresponds to the number of patients in cluster *C* that had an episode, *b* represents the number of patients in the same *C* that did not have an episode, *c* is the number of patients not in *C* that had an episode and *d* the number of patients not in *C* that did not have an episode.

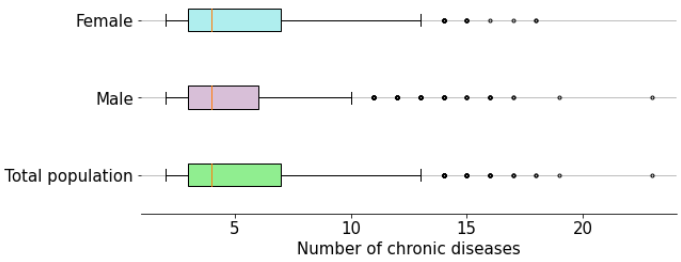
IV. Results & discussion

A. Initial phenotype screening of Dementia patients

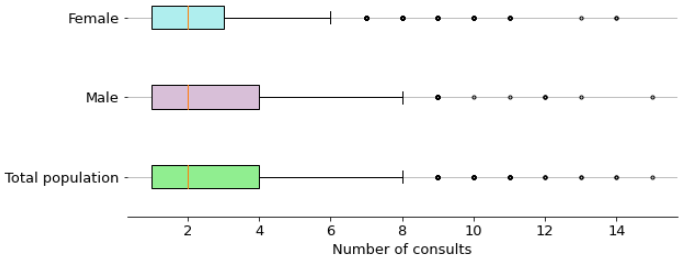
Fig. 1 presents the distributions for Dementia patients age, number of associated chronic diseases and different medical speciality consults attended. With respect to the age spectrum of Dementia patients, presented in Fig. 1 (a), it is possible to see that the distribution is very similar for the whole population, as well as for males and females individually. These patients oscillate in age between 58 and 102 years old, specifically between 61 and 102 for male patients and 60 and



((a)) Age distribution.



((b)) Number of chronic diseases distribution.



((c)) Distribution of the number of different medical speciality consults per patient.

Fig. 1: Distributions of phenotypes that characterize Dementia patients, for the population as a whole and by gender.

100 for females. The first age quartile for all Dementia patients is set to 76 years old, meaning that the majority of Dementia patients (75%) are older than that. It is also possible to observe some points outside of the minimum age value defined by the box plot (58 years old), which represent patients identified as outliers, meaning that these patients fall out of the most common age group. Comparing female and male patients, it is possible to identify the existence of a younger male population with Dementia. Moving on to the distribution of the number of chronic diseases that these Dementia patients suffer from, shown in Fig. 1 (b), it is possible to see that the distribution is identical for the whole population and for females, whilst for male patients there are slight differences. Dementia patients, in general, suffer between two and thirteen chronic diseases, as indicated by the minimum and maximum values of the box plot, being that male patients normally have at most ten. The median number of chronic diseases for this population is set to four, meaning that fifty percent suffer between two and four, while the remaining 50%, when considering female patients, suffer between four and thirteen chronic illnesses and males experience between four and ten. Male patients in this cohort that find themselves displaced from this distribution can have up to twenty-three chronic diseases, while females can admit up to eighteen. Finally, Fig. 1 (c) shows that 50% of Dementia patients, considered as a whole or by gender, have at most two different medical speciality consults in their history. Furthermore, while for female patients the maximum number of different consults is set to six, for male patients this value is set to eight. In the same line of thought, considering the quantiles defined by the box plots, 75% of female patients attended at most three different medical speciality consults, while male patients attended at most four different ones. Regarding the ones that fall out of this distribution, some male patients presented up to fifteen different medical speciality consults in their history, while females had up to fourteen.

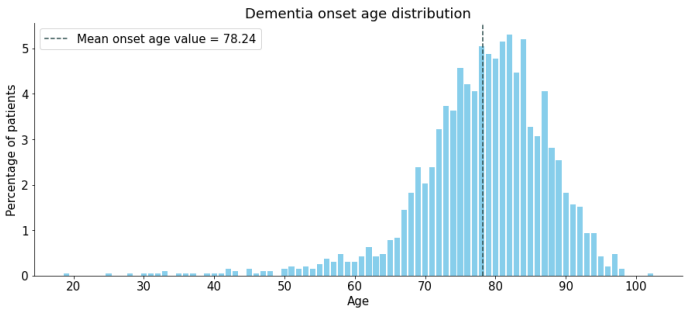


Fig. 2: Dementia onset age distribution in the cohort under study.

Considering that Dementia was identified as presenting a certain heterogeneity in individual onset ages [16], a distribution of these patients age when diagnosed was obtained and can be seen in Fig. 2. Usually, Dementia is a disease which is diagnosed in patients with sixty-five years old or more, being that bellow this age it is considered a premature diagnosis. Cases of patients with onset ages between thirty and sixty-five years old are rare but possible. This early onset Dementia can be a result of post-traumatic experiences, substance abuse issues or genetic reasons. The U.S. Department of Health and Human Services published an article in the National Institute on Ageing journal stating that young-onset Dementia can be a consequence of an inherited mutation in one of three genes [18]. These mutations cause abnormal protein production which leads to the early development of symptoms. Due to the fact it is rare to have a Dementia diagnosis in this age group, physicians do not always look for a Dementia diagnosis in such a young age, which may interfere with the process of early identification of this illness. Furthermore, its symptoms may overlap with those of psychological illnesses, which once again may cause misdiagnosis or delay in the diagnosis of Dementia. Cases of onset in people with less than thirty years-old are very rare and have very few mentions in the literature. In our study cohort, three people with Dementia were diagnosed with less than thirty years old, one at nineteen, another at twenty-five and one with twenty-eight.

Since focusing on patients with multimorbidity, it was important to identify the chronic diseases present in this data set, as well as their prevalence and co-occurrence. There were one hundred and three distinct chronic illnesses identified amongst patients with Dementia, making it important to analyse which

are the most prevailing ones, aiming at understanding which diseases may be more or less related to Dementia. To this end, the top fifteen most common chronic diseases, besides Dementia, in this cohort were identified. Furthermore, since some of these illnesses are very common within the elder population and in order to understand how far Dementia patients are more or less susceptible to suffering from these diseases when compared to patients that suffer from MM in general, the ratio between the fraction of Dementia and multimorbidity patients that suffer from each one of the fifteen mentioned chronic diseases was obtained. Fig. 3 represents the ratio between the fraction of Dementia patients and MM patients in general, that suffer from each one of the top fifteen chronic diseases, which tells us how much more likely a patient with Dementia is to develop a certain disease when compared to the overall MM population.

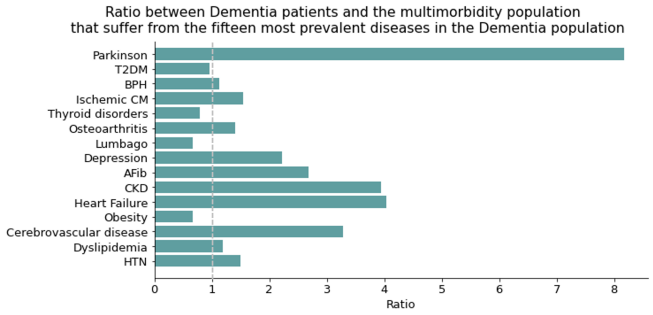


Fig. 3: Incidence of the top fifteen chronic diseases in Dementia patients relative to the MM population.

A ratio around one says that the prevalence of that disease is practically the same whether a patient suffers from Dementia or not, while a ratio below one indicates that Dementia patients are less susceptible of developing that co-morbidity compared to MM patients and a ratio higher than one means the opposite. For instance, the highest ratio is relative to Parkinson's disease, which says that a patient with Dementia is eight times more likely of also suffering from Parkinson than a patient with MM but that does not suffer from Dementia. Additionally, there is a type of Dementia, which is the vascular one, where patients that suffer from a vascular disease have higher probability of developing Dementia symptoms. It is possible to verify this preface by looking at the ratios corresponding to Cerebrovascular disease, Heart Failure, CKD and AFib, which represent risk factor for Vascular Dementia, since these patients are two to four times more probable of suffering from Dementia and one more of these illnesses, when compared to the general MM population. Finally, Depression has also been previously linked to Dementia and with this analysis it is possible to see that it is 2.22 times more probable that Depression co-exists with Dementia than with other disorders in general.

To finalize this initial phenotype screening process, the prevalence of each medical speciality present in the data set was assessed. A population pyramid with this information is presented in Fig. 4. It is interesting to observe that the prevalence of each of the medical speciality consults is similar for both female and male patients. Moreover, being a neurological disease, it is immediate that Neurology consults are the most occurring consults amongst Dementia patients. As it is possible to observe, a great part of the population (approximately 72%) had at least one Neurology consult throughout their hospital activity. Beyond Neurology consults, General and Family Medicine (GFM), Internal Medicine, Anesthesiology and Cardiology consults are very present in these patients history, while Immunoallergology, Hematology, Rheumatology, Nephrology and Dental are the five less attended medical speciality consults.

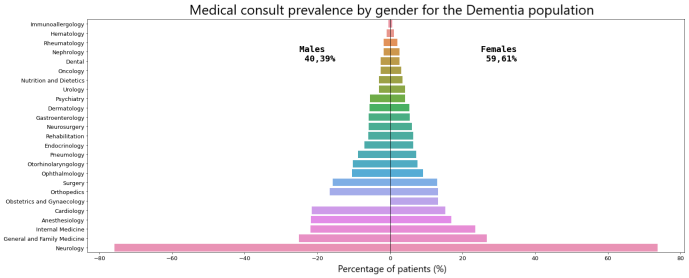


Fig. 4: Percentage of patients, by gender, that attended each medical speciality consult at least once.

B. Clinical pathway analysis - Markov Chains

In this first stage, two transition matrices were initially obtained and visualized in heatmaps, as shown in Fig. 5, for the two mentioned cases. The first one considering consecutive transitions between the same medical speciality consult, as presented in Fig. 5 (a), for which only patients with one consult in their history were filtered, remaining then 1607 Dementia patients, from which 951 are female and 656 male, from the 1865 patients considered for this analysis. The second one does not consider consecutive transitions between the same medical speciality, given by Fig. 5 (b), aggregating consecutive occurrences of the same consult into one and filtering patients who ended up with one consult alone, remaining 1204 patients, 709 females and 495 males, to be considered for the transition matrix formulation.

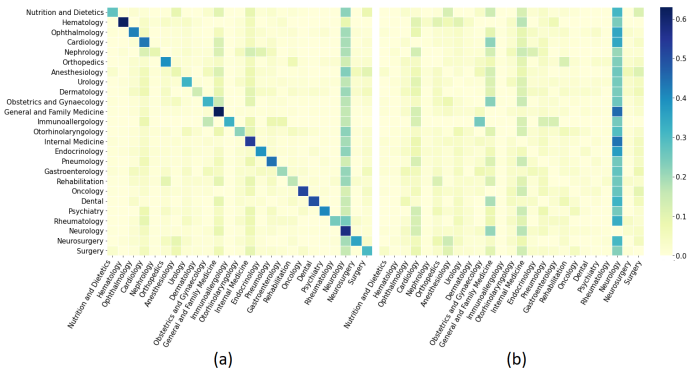


Fig. 5: Heatmaps displaying the transition probabilities of Dementia patients moving between medical speciality consults, when (a) considering and (b) not considering consecutive transitions between the same medical speciality consult.

With respect to the first scenario, shown in Fig. 5 (a), where consecutive transitions between the same medical speciality are considered, it is clear that the most prevailing transitions are between the same consult. This may be indicative of the fact that patients generally have a follow-up consult in the same medical speciality prior to being redirected to a different one. It is also interesting to notice that, since dealing with a neurological illness, independently of the source consult, one of the most probable transitions is to a Neurology appointment. Furthermore, despite not being as clear, it is also possible to observe a slight tendency of Dementia patients of transitioning to Internal Medicine and GFM consults, which goes accordingly to the fact that both these medical specialities are prevailing amongst Dementia patients.

Regarding the second scenario in Fig. 5 (b), where consecutive transitions between the same medical consult are not considered, it is still clear that transitions to Neurology appointments still prevail, no matter what the consult of origin is. Transitions to Internal Medicine, GFM and Cardiology consults can also be prevailing, which makes sense considering the prevalence of these medical specialities in the data set.

C. Clinical pathway analysis - AliClu

1) Preprocessing and parameter optimization: In view of what was mentioned in the previous section, after obtaining the PE sequences for each of the 1865 patients, it was necessary to filter these patients before moving on to the clustering process. Thus, patients with only one medical consult and patients who fall into the 5% over the 95 percentile, are filtered from the data set, remaining 1118 patients with Dementia available for undergoing the sequence alignment and clustering algorithm.

Regarding the parameter optimization process, after going through every phase described in the previous section, we reached several conclusions: (i) the first round of parameter optimization revealed that gap penalties of 0.1 and 0.2 and a temporal penalty of 10 would be the best choices considering the dataset in hands; (ii) Ward's linkage method for the hierarchical clustering process of AliClu was the best choice; (iii) a number of bootstrap samples of 250 is enough, since higher values do not have an influence in the results; (iv) gap penalties outside of the range of values tested in the first round lead to worst clustering outcomes (v) the optimum number of clusters can be very heterogeneous depending on the amount of data, so it should be chosen when considering the entire cohort.

2) Obtaining the final clusters: Given the results of the preliminary optimization, we ran the algorithm for the entire cohort, testing both gap penalty values (*i.e.*, 0.1 and 0.2) which proved to result in good outcomes, setting the temporal penalty to 10, the linkage function to the Ward's method and the number of bootstrap samples to 250 and iterated the number of clusters from 2 to 20. The average silhouette score evolution with the increase of k is presented in Fig. 6. Plots show that the maximum average SS is reached when g is set to 0.1, Tp to 10 and k to 12 clusters, yielding a value of approximately 0.25. Thus, the 12 clusters obtained with these established parameters were chosen to be further analysed.

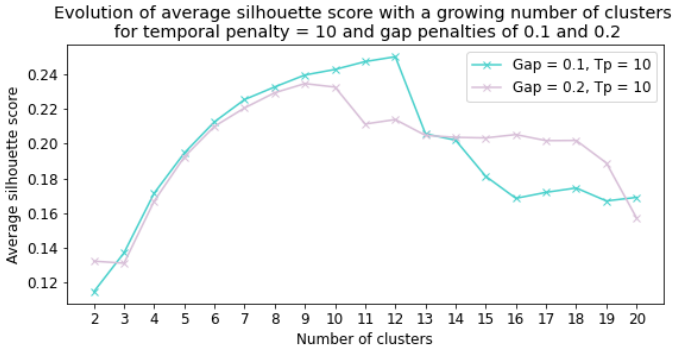


Fig. 6: Evolution of the average SS with the number of clusters formed from $k = 2$ to $k = 20$, for a Tp of 10 and gap penalties 0.1 and 0.2

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

TABLE I: Number of elements and prevailing medical speciality consult within each cluster.

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

Moreover, in order to quantitatively analyse the content present in each one of the twelve clusters formed with (g = 0.1, Tp = 10, k = 12), it is essential to look at Fig. 7, where the SS for each sample in the data set, as well as the average SSs for each cluster and across all samples is represented. Keeping in mind that the SS measures how close a sample is to elements belonging to the same cluster, in comparison with how close or far that same sample is from other elements in different clusters, it is possible to verify that cluster 9 is composed by outliers. This conclusion is reachable due to the fact that the silhouette scores of all samples in this cluster have very low values, being that the majority of the elements even have negative scores. Furthermore, it is also possible to identify outliers in several other clusters, considering that some elements of clusters 2, 6, 8, 9, 10 and 12 possess negative silhouette scores.

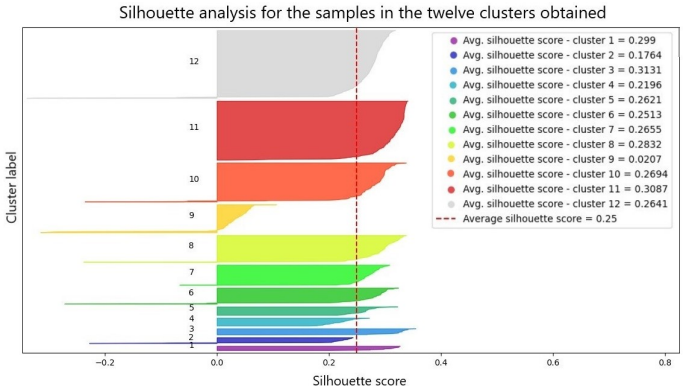


Fig. 7: Silhouette scores for each sample present in each of the obtained twelve clusters for $g = 0.1$ and $Tp = 10$, average SS per cluster and across all samples in the data set.

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

Fig. 8: Directed graphs showing the transitions between different medical consults within clusters (a) 3, (b) 5, (c) 11 and (d) 12, where Orthopedics, Ob-Gyn, Neurology and GFM consults prevail, respectively. Bigger and darker colored nodes indicate higher occurrence of that consult in the cluster, while wider edges indicate more patients underwent that transition at least once. The nodes are displayed in a circular manner, ordered by the prevalence of the type of appointment.

In order to obtain a more visual look at the clusters formed, a directed graph was developed for each cluster, where nodes represent medical speciality consults existing within each cluster and edges serve as the number of patients who underwent a certain transition at least once. As an example of the obtained graphs, Fig. 8 represents the activity within clusters 3, 5, 11 and 12 where it is clear that Orthopedics, Obstetrics and Gynaecology, Neurology and GFM, respectively, are the dominant consults. Beyond the strong transition wise relation between the first consult from each cluster with Neurology consults, it is possible to identify certain patterns, in come clusters more than others. For instance, Fig. 8 (d), representing the cluster with GFM prevalence, is an heterogeneous one, since it is not possible to identify prominent transitions as it is possible to pinpoint in others.

D. Phenotype screening per cluster

An analysis of gender, age and chronic diseases within each cluster obtained was done. The gender analysis revealed that the female/male ratio per cluster is not significant, except for the Ob-Gyn and anesthesiology clusters, where the discrepancy between genders is not due to the discrepancy in the whole data set. The age distribution indicates that the Ob-Gyn cluster is the one that diverges more from the original distribution, being formed by younger patients. Regarding the incident chronic diseases in each cluster, the top four were assessed. Results can be seen in Fig. 9.

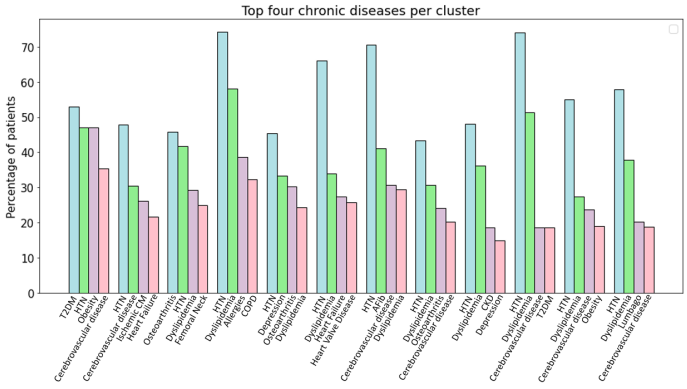


Fig. 9: Incidence of the four most prevalent chronic diseases per cluster. Each set of four bars represents the incidence in each one of the sub-groups, which are sorted, from left to right, in ascending order. Blue bars indicate the chronic illness suffered by most patients in the cluster, while pink bars demonstrate which one is suffered by the least amount of individuals. Green and purple bars represent the second and third most prevailing chronic diseases in the subset, respectively.

Paying attention to the blue bars, representing the most prevailing chronic disease in each cluster, besides Dementia, it is possible to see that Hypertension (HTN) is the disease that dominates almost all sub-groups (ten out of twelve), with the exception of the first and third ones, where more patients suffer from Type 2 Diabetes (T2DM) and Osteoarthritis, respectively. Besides HTN, Dyslipidemia is also a very common disease

amongst these patients. As it is possible to see, it is present in the top four chronic diseases of ten of the total number of clusters, mostly as the second most incident, with exception of numbers three, five and seven, where it is the third or fourth most prevalent one. Moreover, a majority of even out of twelve clusters are formed by a considerate amount of patients that suffer from Cerebrovascular disease. However, it was already identified in the initial phenotype screening that these three mentioned diseases have high incidence in Dementia patients.

It is possible to identify specific chronic disease patterns that can be associated to the prevalence of certain medical specialities in each sub group, namely, within the Endocrinology, Orthopedics, Pneumology, Ob-Gyn, Surgery and Cardiology clusters.

E. Medication analysis

The preliminary analysis done to identify the fraction of patients who did not have any registered medication, as well as the average number of prescriptions and average number of different medication taken per patient, indicated that:(i) 16.9% of the 1118 Dementia patients did not have any registered prescriptions, (ii) 25 is the average number of prescriptions per patient and (iii) 11 is the number of different medication intake per patient.

The next step was identifying the top 5 drugs prescribed to patients per cluster obtained. Results are presented in Fig. **??**. It is possible to establish some associations between the most prevalent medications in each cluster with the most incident medical consults and chronic diseases. For instance, the first cluster, which is the Endocrinology one, has prevalence of medication related to the treatment of diabetes, which makes sense both with the most incident medical speciality, as well as with the high prevalence of type 2 diabetes in patients that form this subgroup. The Ob-Gyn cluster has estriol as most common medicine, which is a drug given to female patients, normally to treat symptoms related to menopause. Two types of vitamins are also recurrent, namely cholecalciferol and calcium carbonate. The seventh cluster has rosuvastatin as the most common medicine and furosemide as the fourth most common, which goes accordingly to the fact that these are used to prevent cardiac problems, as mentioned in the previous item. Also included in the top five medications is bisoprolol, which is used to treat high blood pressure and dabigatran etexilate, which is prescribed to treat and avoid blood clots, which all go accordingly to the fact that this group gathers cardiac patients.

TABLE II: Prevalence of twenty classes of medications in each cluster and in the whole data set.

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

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

F. Hospital admission and emergency analysis

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

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

An initial analysis was done to identify the fraction of Dementia and MM patients with hospital admissions (HA) and emergency episodes (EE), as well as the average number of each occurrence in the time frame being considered, which results can be observed in Table III.

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

Cluster	HA Odds Ratio	EE Odds Ratio
Cluster 1	0.81	0.94
Cluster 2	2.21	2.13
Cluster 3	0.57	1.00
Cluster 4	0.84	1.90
Cluster 5	0.85	0.62
Cluster 6	2.05	1.61
Cluster 7	1.47	1.24
Cluster 8	1.40	0.89
Cluster 9	0.60	0.35
Cluster 10	1.52	1.45
Cluster 11	1.00	0.69
Cluster 12	0.60	1.99

Moving on to the assessment on how and if each cluster is associated to a higher or lower chance of being admitted to the hospital or to the emergency room, the odds ratio associated with each cluster and each of the events are presented in Table IV. Focusing on the hospital admissions table and keeping in mind that an OR smaller than one means that belonging to a certain cluster leads to lower odds, we can see that clusters one, three, four, five, nine and twelve are in this situation. These are the subgroups with prevalence of Endocrinology, Orthopedics, Pneumology, Ob-Gyn, GFM and the one that gathers the *outliers*, indicating that these patients are less probable of being admitted, when compared to others. On the other hand, subsets of Nutrition and Dietetics, Surgery,

Cardiology, Anesthesiology and Internal Medicine patients have an OR higher than one. This indicates that Dementia patients belonging to these clusters are associated to higher odds of being admitted. Lastly, cluster eleven, which is the one that gathers Neurology patients is the only one which does not influence the odds of the outcome, meaning that by belonging to this cluster, a patient is not more or less probable of being admitted.

Regarding the ORs regarding emergency episodes, we can see that belonging to the Endocrinology, Ob-Gyn, Anesthesiology, Neurology and the *outliers* subgroups is associated to lower odds of having an emergency episode. On the other hand, being part of Nutrition and Dietetics, Pneumology, Surgery, Cardiology, Internal Medicine and GFM clusters is associated to higher odds of emergencies. The Orthopedics subset has an OR of one, meaning that this exposure does not affect the emergency outcome.

V. Conclusions

In the present study we focused on the analysis of clinical pathways, by exploring available data of patients with Dementia from HLL, resorting to Markov Chains and AliClu.

Furthermore, we performed a phenotype evaluation on the chosen cohort, both prior to the clinical pathway analysis, to better understand the patients under study, as well as subsequently to obtaining the resulting clusters. By resorting to directed graphs, it was possible to visualize the most prevalent clinical pathways regarding Dementia patients, per cluster formed, from which conclusions were drawn concerning most visited medical speciality consults, as well as most common transitions.

We were able to expose heterogeneity amongst patients, as well as activity patterns. The methods used in this work allowed to pinpoint the prevailing attended consults within Dementia patients, as well as to identify patterns when considering consult transitions. When coupling this analysis to a phenotypic screening we obtain an auxiliary tool to provide an early overview of the patients' most probable pathway patterns, allowing health providers to align and optimize their offer to their patients needs, for instance, when dealing with other diseases, with a certain age or gender group, or even if wanting to investigate different temporal events. Big data analysis of electronic medical records can create tools to support health care providers and growing knowledge on how to deal with heterogeneous sources of health care data opens doors to new possibilities and advantages to the healthcare sector.

Future work which we believe would improve the performance of the methods used include the use of natural language processing algorithms to extract information from clinical notes. Additionally, in order to further validate the patterns that were identified among Dementia patients, it would be advantageous to test these methods in other disease datasets.

References

[1] M. Rijken et al. *How to improve care for people with multimorbidity in Europe?* European Observatory on Health Systems and Policies - Policy Brief no. 23. 2017.

[2] A. Shinozaki. "Electronic Medical Records and Machine Learning in Approaches to Drug Development." In: *Artificial Intelligence In Oncology Drug Discovery And Development*. (2020). DOI: 10.5772/intechopen.92613.

[3] Juan M Banda et al. "Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models." In: *Annual review of biomedical data science* 1 (2018), pp. 53–68. DOI: 10.1146/annurev-biodatasci-080917-013315.

[4] C. Violan et al. "Prevalence, Determinants and Patterns of Multimorbidity in Primary Care: A Systematic Review of Observational Studies." In: *Plos ONE, 9(7)* (2014). DOI: 10.1371/journal.pone.0102149.

[5] R. Navickas et al. "Multimorbidity: What Do We Know? What Should We Do?" In: *Journal Of Comorbidity, 6(1)* (2016), pp. 4–11. DOI: 10.15256/joc.2016.6.72.

[6] A. Hassaine, D. Canoy, and J.R.A. Solares et al. "Learning multimorbidity patterns from electronic health records using Non-negative Matrix Factorisation". In: *Journal of Biomedical Informatics* (2020). DOI: https://doi.org/10.1016/j.jbi.2020.103606.

[7] O. Ben-Assuli, R. Padman, and I. Shabtai. "Exploring trajectories of emergency department visits using a laboratory-based indicator of serious illness". In: *Health Informatics Journal* 26 (2019), pp. 205–217. DOI: 10.1177/1460458218824751.

[8] Haytham Elghazel et al. "Clinical pathway analysis using graph-based approach and Markov models". In: *2007 2nd International Conference on Digital Information Management*. Vol. 1. 2007, pp. 279–284. DOI: 10.1109/ICDIM.2007.4444236.

[9] Renato Cesar Sato and Désirée Moraes Zouain. "Markov Models in health care". In: *Einstein (Sao Paulo)* (2010).

[10] Francesco Folino and Clara Pizzuti. "Combining Markov Models and Association Analysis for Disease Prediction". In: *Information Technology in Bio- and Medical Informatics*. Ed. by Christian Böhm et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 39–52. ISBN: 978-3-642-23208-4.

[11] Sula Windgassen et al. "The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome". In: *Journal of Mental Health* 27.2 (2018). PMID: 29447026, pp. 94–96. DOI: 10.1080/09638237.2018.1437615.

[12] Kishan Rama et al. "AliClu - Temporal sequence alignment for clustering longitudinal clinical data". In: *BMC Medical Informatics and Decision Making* 19 (2019). ISSN: 1472-6947.

[13] Alexia Giannoula et al. "Identifying Temporal Patterns In Patient Disease Trajectories Using Dynamic Time Warping: A Population-Based Study." In: *Scientific Reports* 8 (2018). DOI: 10.1038/s41598-018-22578-1..

[14] World Health Organization. *Dementia*. https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed: 2021-09-28.

[15] J. Ryan et al. "Phenotypic Heterogeneity in Dementia: A Challenge for Epidemiology and Biomarker Studies". In: *Front Public Health* 6 (2018). DOI: 10.3389/fpubh.2018.00181.

[16] Serena Verdi et al. "Beyond the average patient: how neuroimaging models can address heterogeneity in dementia". In: *Brain* (Apr. 2021). awab165. ISSN: 0006-8950. DOI: 10.1093/brain/awab165.

[17] Magdalena Szumilas. "Explaining odds ratios." In: *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 19 (2010), pp. 227–229. ISSN: 2293-6122.

[18] *Alzheimer's Disease Genetics Fact Sheet*. 2019. URL: https://www.nia.nih.gov/health/alzheimers-disease-genetics-fact-sheet.