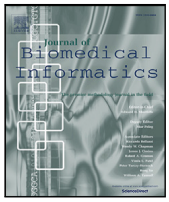
Journal of Biomedical Informatics 140 (2023) 104328

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Using Markov chains and temporal alignment to identify clinical patterns in Dementia

Luísa Marote Costa [d,a], João Colaço [b], Alexandra M. Carvalho [c,a], Susana Vinga [d,a,*], Andreia Sofia Teixeira [e,d,f]

[a] Instituto Superior Técnico, Avenida Rovisco Pais, 1, Lisbon, 1049-001, Portugal
[b] Hospital da Luz Lisboa, Av. Lusiada 100, Lisbon, 1500-650, Portugal
[c] Instituto de Telecomunicações, Av. Rovisco Pais 1, Lisbon, 1049-001, Portugal
[d] INESC-ID, Rua Alves Redol 9, Lisbon, 1000-029, Portugal
[e] LASIGE, Departamento de Informatica, Faculdade de Ciencias, Universidade de Lisboa, Campo Grande 016, Lisboa, 1749-016, Portugal
[f] Hospital da Luz Learning Health, Luz Saúde, Portugal

ARTICLE INFO

MSC:
0000
1111

Keywords:
Multimorbidity
Dementia
Markov chains
Temporal sequence alignment
Clustering
Electronic medical records

ABSTRACT

In the healthcare sector, resorting to big data and advanced analytics is a great advantage when dealing with complex groups of patients in terms of comorbidities, representing a significant step towards personalized targeting. In this work, we focus on understanding key features and clinical pathways of patients with multimorbidity suffering from Dementia. This disease can result from many heterogeneous factors, potentially becoming more prevalent as the population ages. We present a set of methods that allow us to identify medical appointment patterns within a cohort of 1924 patients followed from January 2007 to August 2021 in Hospital da Luz (Lisbon), and to stratify patients into subgroups that exhibit similar patterns of interaction. With Markov Chains, we are able to identify the most prevailing medical appointments attended by Dementia patients, as well as recurring transitions between these. To perform patient stratification, we applied AliClu, a temporal sequence alignment algorithm for clustering longitudinal clinical data, which allowed us to successfully identify patient subgroups with similar medical appointment activity. A feature analysis per cluster obtained allows the identification of distinct patterns and characteristics. This pipeline provides a tool to identify prevailing clinical pathways of medical appointments within the dataset, as well as the most common transitions between medical specialities within Dementia patients. This methodology, alongside demographic and clinical data, has the potential to provide early signalling of the most likely clinical pathways and serve as a support tool for health providers in deciding the best course of treatment, considering a patient as a whole.

1. 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 complex conditions, suffering from multiple comorbidities. Multimorbidity can be defined as more than one chronic or long-term disease. Despite its prevalence in the population, especially in the elderly 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, while offering 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.

* Corresponding author at: INESC-ID, Rua Alves Redol 9, Lisbon, 1000-029, Portugal.
E-mail address: susanavinga@tecnico.ulisboa.pt (S. Vinga).

https://doi.org/10.1016/j.jbi.2023.104328
Received 16 May 2022; Received in revised form 23 February 2023; Accepted 6 March 2023
Available online 14 March 2023

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

The heterogeneity of Dementia patients 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 aetiology and pathophysiology [2]. 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 [3]. It has been evidenced that there is a high incidence of comorbid medical conditions among patients with Dementia [4], being estimated that these can suffer from two to eight additional chronic illnesses [5]. Incidence of certain co-existing diseases in the population with Dementia may even aggravate the patient's condition, for instance, as stated by [4], type 2 diabetes may accelerate the cognitive decline of a patient that suffers from Dementia. This high incidence of additional chronic diseases associated with these patients brings many challenges to the healthcare system since, for instance, under-diagnosis and treatment of Dementia may result from an accelerated advancement towards deteriorated cognitive and functional states due to co-existing illnesses [4,5]. Additionally, these patients show high utilization of health services, representing a compelling portion of costs associated with healthcare concerning the elderly population, being that the prevalence of comorbidities aggravates this usage, increasing hospital stay, healthcare costs and mortality rates for hospitalized patients [5]. Given all these challenges associated with the incidence of Dementia in the population, together with other chronic conditions, it is important to move towards a better understanding of certain relationships between co-existing illnesses in these patients, shifting the focus to treating these patients from a global perspective.

Our main goal is to create a pipeline of existing methods that allow us to analyse and identify patterns within a complex cohort of patients suffering from multimorbidity in which Dementia is included. In this work, we show a set of methods which allows finding recurrent patterns of medical appointments within the entire cohort, as well as stratifying 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 clinical data has the potential to provide early signalling of the most likely clinical pathways. Stratifying the patients based on their activity allows the detection of 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-centred care, considering a patient as a whole and not focusing only on a certain problem.

We use AliClu [6], a temporal sequence alignment algorithm, to cluster longitudinal clinical data and stratify patients accordingly. We conduct a feature analysis for each cluster, which allows us to identify distinct patterns and characteristics. The proposed pipeline provides a tool for identifying prevailing clinical pathways of medical appointments, as well as the most common transitions between medical specialities for dementia patients. This methodology, along with demographic and clinical data, has the potential to provide early signalling of the most likely clinical pathways and can serve as a support tool for healthcare providers in deciding the best course of treatment for patients, taking into account their overall health. In the present, we are deepening the external validation of the work with the clinicians by studying the impact of reorganizing dementia appointments at the hospital for some of these patient groups, so that they can receive them more quickly.

2. Materials and methods

In this study, a pipeline was developed to identify characteristics and patterns within Dementia patients that suffer from multimorbidity, regarding their features and clinical pathways. The pipeline includes the following steps: (i) initial variable screening to characterize the dataset; (ii) creation of transition matrices to identify the most common medical appointments activity; (iii) clustering algorithm based on medical appointments pathway; (iv) characterization, visualization and variable screening of the obtained clusters, including age, gender, chronic diseases, medication, hospitalization and emergency analysis.

2.1. Available data and initial feature analysis

In order to fulfil our goals and motivation, data was collected from Hospital da Luz Lisboa (HLL), from January 2007 to August 2021. Collected data included 302,709 patients, 63,786 of them suffering from multimorbidity. Among these, 1924 (1147 female and 777 male) were identified as having Dementia. Data concerning these patients' age, gender, and chronic diseases, as well as information on 20,033 medical appointments, attended, was gathered. An initial approach involved characterization of the data set in terms of the collected features, 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's 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 appointments. 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. The flowchart in Fig. 1 shows the process of inclusion and exclusion of patients from the initial data set until obtaining the final patient cohort, used in our study. To sum up, the Dementia patients were identified by ICD9 codes and keywords related to the disease. The first filtering step from the initial patient dataset was to gather patients with multimorbidity. This means, only patients with more than one diagnosed chronic disease moved to the next step. Within this pool of multimorbidity patients, the next step was to filter patients diagnosed with Dementia. The final inclusion criterion was based on the availability of medical appointment records, within the identified Dementia cohort.

Within the 20,033 medical appointments records gathered for these patients, twenty-five different medical specialities were identified. These are listed in Table 1, with the respective event label used for the clinical pathway analysis assigned.

Furthermore, in order to accomplish the medication and hospital admissions analysis, data on 35,263 prescriptions given to Dementia patients, 2078 hospital admissions (HA) and 9991 emergency episodes (EE) were 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.

2.2. Clinical pathway analysis

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

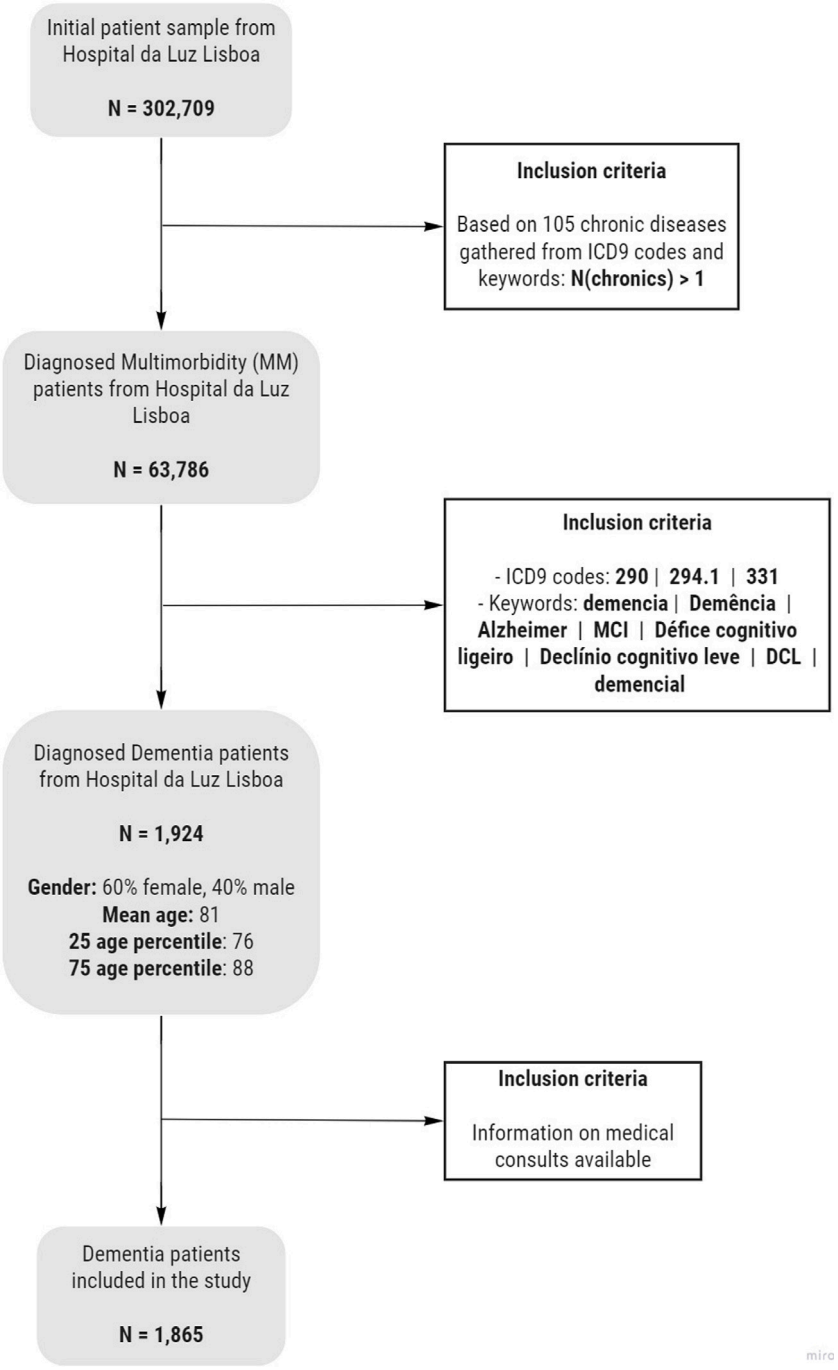


Fig. 1. Flow chart specifying inclusion criteria until obtaining the final multimorbidity Dementia patient cohort used in the study.

is by analysing their clinical pathways [7]. Two distinct approaches were used to this end: (i) Markov chains and (ii) a patient stratification algorithm based on temporal sequence alignment, named AliClu [6]. Markov Chains were used to study the most common transitions between medical specialities within the entire Dementia cohort. On the other hand, AliClu was used to stratify these patients, grouping them by similarity based on their clinical pathways patterns.

2.2.1. Markov chains

By resorting to a Markov Chain model of order one, we can explore a probabilistic model that is able to identify patterns in clinical pathways at the population level, as it has been done in [8]. Markov Chains 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 appointments.

Through estimation of this Markov Chain, it is possible to determine the most prevalent transitions between medical appointments in a population. This is accomplished by formulating a transition matrix (TM), which shows the transition probabilities between two states (i.e., medical appointments). Given a square matrix of all possible medical specialities, we calculate the conditional probabilities of moving to a second speciality appointment (j), given the previous one (i). This is achieved by dividing the number of times that each transition occurs by the prevalence of the original medical speciality appointment, filling the transition matrix as follows:

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

Table 1
Medical specialities considered for patient activity analysis, along with the corresponding event label assigned by the algorithm.

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

Two different approaches are used 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 patterns without considering follow-up appointments in the same speciality.

2.2.2. AliClu algorithm

An alternative approach to address the heterogeneity amongst clinical populations and how to target it is through patient stratification [10]. Dynamic time warping (DTW) has been used to find temporal patterns in patient disease trajectories [11]. However, to our knowledge, AliClu [6] is the only method able to address mixed longitudinal data considering a sequence of events (e.g., medical appointments or treatments) and the time between events. AliClu is an algorithm that combines temporal sequence alignment and hierarchical clustering; it was developed and applied to a set of Rheumatology patients to stratify them based on their medication switches throughout time. 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 performed, resorting to pairwise scores obtained during the alignment process. In this work, we adapt this algorithm to our data. First, the available data from HLL concerning the medical appointments activity of patients with Dementia was preprocessed, converting 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.

Preprocessing

Converting the medical appointments data from panel data format to the correct input temporal sequences, named prefix-encoded (PE) sequences, involved the following steps: (i) each event, in this case, a medical appointment, is assigned a label; (ii) the time elapsed between consecutive events is calculated and fixed in the sequence in days; and (iii) 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 appointment 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 appointments superior to a certain threshold. This threshold is given by the 95 percentile regarding the number of different medical appointments 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 appointment activity.

Parameter optimization

AliClu was designed to return one set of clusters for the best combination of gap penalty (g), and the number of clusters (k), being the temporal penalty (T_p), established at 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 the silhouette score (SS) clustering index, decide on the set of parameters that would 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.

Obtaining the final clusters

The AliClu parameter optimization was done as follows. The gap penalty g was varied from −0.5 to 0.5 with a step increase of 0.1, while the temporal penalty T_p was tested with values of 1, 2, 5, 7, and 10. Additionally, the number of clusters k varied between 2 and 20. When performing parameter optimization, we concluded that negative gap penalties g resulted in low average SS values. Regarding the results for positive g, it became clear that the gap penalty of 0.1 performed better in terms of average SS as well as by observation of the clusters. Regarding the influence of the temporal penalty T_p in the results, there was an apparent increase in the values of average SS proportional to an increase in T_p. To sum up, the best results were obtained with gap penalties of $g = 0.1$, while the temporal penalty was set to 10 ($T_p = 10$) and the number of clusters 12 ($k = 12$). These parameters were set to run the algorithm for the Dementia data set.

2.3. Clustering and feature analysis

After setting the optimal AliClu parameters, a visual inspection of the optimum groups was done to confirm that the algorithm was indeed grouping patients that presented a similar pathway regarding appointment attendance. Moreover, besides the average SS calculated across all samples, the mean Silhouette was also calculated within each cluster to assess the similarity of the elements within a particular cluster when compared to elements allocated to different groups.

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

2.4. 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 from January 2007 until August 2021. Then, the average number of each occurrence per patient was also assessed. These statistics were then compared with

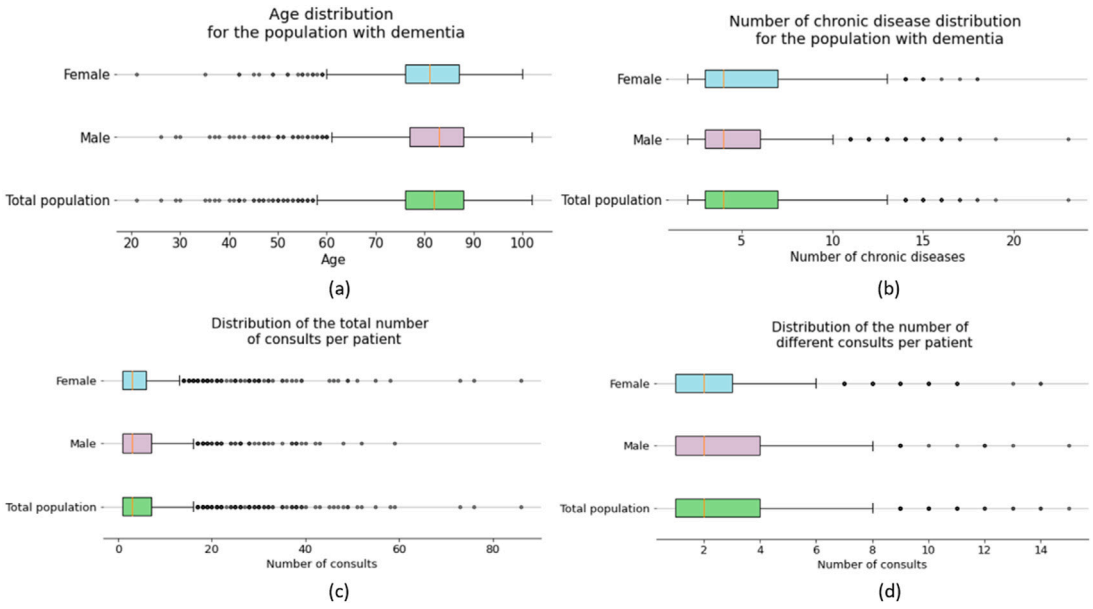


Fig. 2. Distributions of features that characterize Dementia patients, for the whole population and by gender.

the same statistics but for the whole cohort of multimorbidity patients, in order to assess the higher or lower event probability of Dementia patients, when compared to a control cohort. 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, the odds of presenting a hospital admission or emergency episode per cluster were assessed.

3. Results and discussion

3.1. Initial characterization and variable screening

Results of the initial characterization are presented in Fig. 2. With respect to the age spectrum of Dementia patients, presented in Fig. 2(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, approximately.

Moving on to the distribution of the number of chronic diseases that these Dementia patients suffer from, shown in Fig. 2(b), it is possible to see that, in general, male patients suffer from fewer co-existing chronic diseases, when compared to female patients.

Finally, regarding the distributions of the number of appointments attended by Dementia patients, two distinct analyses were carried out. The first one considers all medical appointments attended (Fig. 2(c)), while the second one only takes into consideration the number of different medical speciality appointments attended by these patients (Fig. 2(d)). Focusing on Fig. 2(c), we see that female patients have a slightly lower appointment attendance when compared to males. However, we can see that 50% of both populations visit the hospital for a medical appointment a maximum of three times. Finally, Fig. 2(d) shows that 50% of Dementia patients, considered as a whole or by gender, have at most two different medical speciality appointments in their history. Female patients, in general, present less variance when considering medical appointment attendance since the distribution of the number of different visits attended is not as wide when compared to male patients.

Since focusing on patients with multimorbidity, it was important to identify the co-existing chronic diseases present in this data set of Dementia patients, 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. Fig. 3 represents the incidence of the fifteen most common comorbidities within this population, relative to their incidence in the general multimorbidity population. The ratios obtained indicate how much a patient with Dementia is more or less predisposed to suffer from a certain disease when compared to a multimorbidity patient that does not necessarily suffer from this disease. 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. As it is possible to see, Hypertension (HTN), Dyslipidemia, Cerebrovascular disease, Obesity, Heart Failure, Chronic Kidney Disease (CKD), Atrial Fibrillation (AFib), Depression, Lumbago, Osteoarthritis, Thyroid disorders, Ischemic Cardiomyopathy (CM), Benign Prostatic Hyperplasia (BPH), Type 2 diabetes (T2DM) and Parkinson's disease represent the top fifteen incident illnesses in the Dementia population.

HTN, dyslipidemia, cerebrovascular disease, obesity, heart failure, AFib, Ischemic CM and T2DM are all comorbidities that represent risk factors for Vascular Dementia, since they represent diseases that affect blood vessels, leading to poor brain irrigation, which leads to Dementia states. On the other hand, osteoarthritis and lumbago are illnesses that generate pain, which may cause difficulty to concentrate and to perform cognitively, leading a patient to loose certain basic functioning capabilities, hence, influencing Dementia states. Regarding depression, it may be dubious, since a depression diagnosis can be many times confused with an early Dementia state that goes undetected. This happens due to the fact that many Dementia states begin with behavioural changes, where patients feel and look debilitated, very much like depression states; however, it may be an early signal for Dementia.

The highest ratio is relative to Parkinson's disease, which says that a patient with Dementia is eight times more likely to also suffer from Parkinson's than a patient with MM, but that does not suffer from Dementia. Moreover, despite the fact that certain risk factors, such as obesity, can be associated with cognitive decline, this analysis demonstrates that Dementia patients, at least the ones that are part of this cohort, are less associated with this risk factor than the MM population. The same is observed for lumbago and thyroid diseases. Despite belonging to the top fifteen most incident chronic conditions in

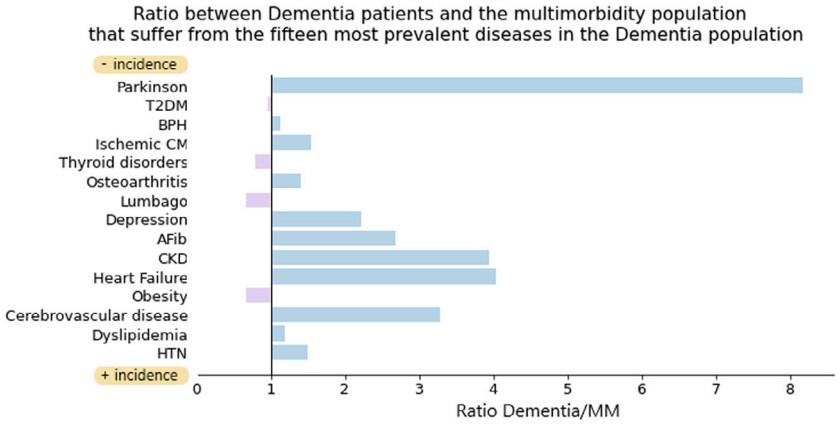


Fig. 3. Incidence of the top fifteen chronic diseases in Dementia patients relative to the MM population, where HTN is the disease with the most incidence of the fifteen and Parkinson the one with less. The vertical line represents the unity threshold from which Dementia patients are more susceptible than MM patients of suffering from one of the top fifteen diseases. A ratio of around one indicates that the prevalence of a disease is nearly the same whether a patient has Dementia or not. A ratio below one suggests that Dementia patients are less likely to develop that particular comorbidity compared to patients without Dementia, while a ratio above one implies the opposite. The highest ratio is relative to Parkinson's disease, which says that a patient with Dementia is eight times more likely to also suffer from Parkinson's than a patient with MM, but that does not suffer from Dementia.

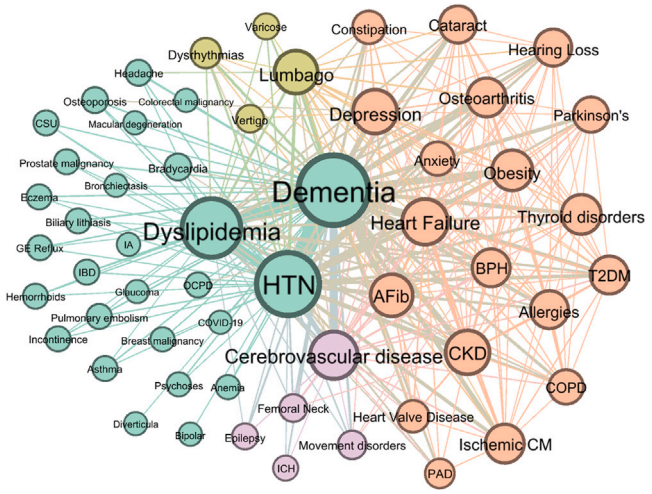


Fig. 4. Graph of chronic disease co-occurrence in Dementia patients, considering only the pair of illnesses that have incidence in more than one percent of the population. Bigger nodes indicate a higher incidence of the chronic disease in the data set and wider edges indicate more co-occurrence of that pair of diseases.

the Dementia population, their prevalence is not necessarily associated with this disease. Lumbago, described as acute lower back pain, and thyroid disorders have proven to have a high incidence in the elder population [12,13]. The fact that Fig. 2 shows a ratio lower than 1 for these diseases means that these diseases have a higher incidence in the MM general population compared to the MM-Dementia population. Hence, we can say that Lumbago and thyroid diseases, despite being in the top 15 incident chronics in this dementia cohort, are also in the general MM population collected from HLuz.

Shifting the focus to the chronic disease co-occurrence graphs obtained for the Dementia cohort under study, represented in Fig. 4, it is important to point out that bigger nodes indicate a higher prevalence of that chronic illness within Dementia patients, while wider edges indicate more co-occurrence of a pair of diseases. The co-occurrence graph obtained for all Dementia patients was filtered at one percent, meaning that what is observable in Fig. 4 is only relative to diseases that co-occur in more than one percent of the population. This filter was applied for relevance in the visualization of co-occurrences, eliminating possible noisy edges.

It is clear from the edges connecting Dementia to other chronic illnesses in the three graphs, that the ones more often in co-occurrence with Dementia are HTN and dyslipidemia. The four distinct disease sub-groups in the graph were obtained by resorting to the modularity property, which is a measure of a network's or graph's structure, assessing the degree to which these can be divided into separate communities that have higher interaction between them when compared to others [14]. It is interesting to see that the majority of the top fifteen previously identified chronic diseases, presented in Fig. 3, are highly interconnected in the orange sub-group. The top two chronic conditions, besides Dementia, which are HTN and dyslipidemia belong to the same sub-group, together with the remaining diseases. In addition, it is possible to observe that cerebrovascular disease and lumbago, despite being in the top fifteen comorbidities of Dementia patients, do not belong to the same community as the remaining ones. Lumbago was identified as being more interconnected with vertigo, varicose and dysrhythmias, while cerebrovascular disease is more related to *Ichthyophthirius multifiliis* (ICH), femoral neck pathologies, epilepsy and movement disorders.

3.2. Clinical pathway analysis

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

First of all, it is important to point out the fact that within the 1924 patients with Dementia present in the study cohort, 59 of them did not have information on their medical appointments. 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.

3.2.1. 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:

- Considering consecutive transitions between the same medical speciality appointment, as presented in Fig. 5(a), for which only patients with one appointment in their history were filtered, remaining 1607 Dementia patients, from which 951 are female and 656 male, from the 1865 patients considered for this analysis.
- Not considering consecutive transitions between the same medical speciality, given by Fig. 5(b), aggregating consecutive occurrences of the same appointment into one and filtering patients who ended up with one appointment alone, remaining 1204 patients, 709 females and 495 males, to be considered for the transition matrix formulation.

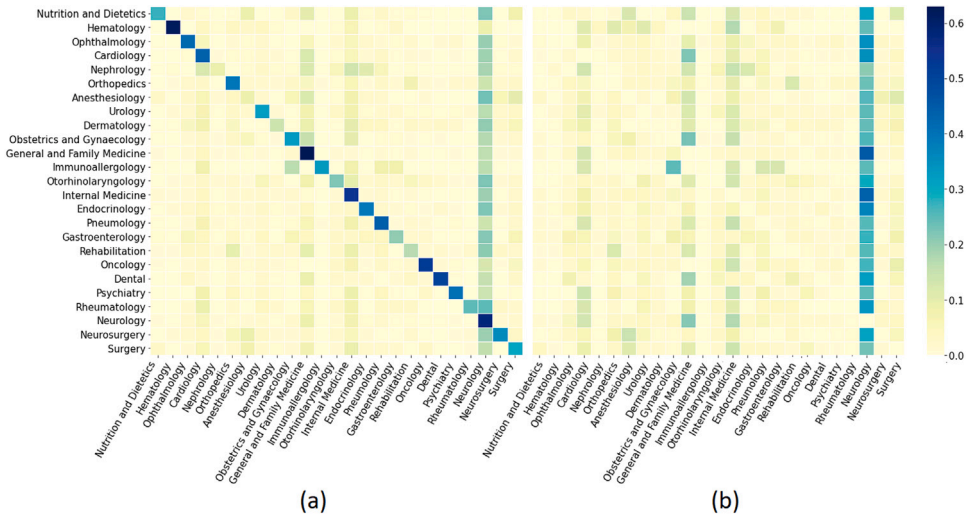


Fig. 5. Heatmaps displaying the transitions between different medical speciality appointments for the whole population with Dementia when both considering (a) and not considering (b) consecutive transitions between the same appointment. Each entrance of the transition matrix is the probability of moving to the column event, knowing that the previous event is represented by the row of the matrix. Also, a transition matrix representing a Markov chain is a stochastic matrix, meaning that each row represents the most probable transitions when the origin is the appointment indicated by the horizontal entrance.

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 appointment. This may be indicative of the fact that patients generally have a follow-up appointment 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 appointment, 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 appointments, which goes accordingly to the fact that both these medical specialities prevail amongst Dementia patients.

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

3.2.2. Obtaining patient clusters with AliClu

The preprocessing step of obtaining the input temporal sequences (PE sequences) of medical appointments for the AliClu algorithm resulted in 1118 patients undergoing the clustering process. A PE sequence was obtained for each of the 1865 patients that had information on medical appointments. Since the focus was to analyse patient activity on multiple fronts, the 661 who presented only one medical appointment in their history were filtered out of the data. Furthermore, in order to avoid the presence of *outliers* in the clustering process as much as possible, patients who had a number of medical appointments higher than the 95 percentile were also filtered out. This threshold was set to twenty-one appointments; hence, the 86 who presented twenty-two or more medical appointments in their history were also filtered out, remaining then 1118 Dementia patients to stratify, based on medical appointment activity.

Subsequently to going through the necessary preprocessing steps to prepare the data for the AliClu algorithm, a parameter optimization process was put in motion, having reached an optimal gap penalty of 0.1, a temporal penalty of 10 and a total of 12 clusters. Furthermore, it was verified that the linkage function for the hierarchical clustering process that led to the best data partitions was, as hypothesized, Ward's method. A number of bootstrap samples of 250 also proved to be the most assertive choice.

Table 2
Number of elements and prevailing medical speciality appointment within each cluster.

Cluster label	Prevailing medical speciality appointment	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

It was curious to notice that the algorithm grouped the patients mainly by the first appointment registered in their medical history. Hence, patients assigned to the same cluster begin their activity in the same medical speciality appointment. Table 2 shows the medical speciality indicating the start of the patients' pathways within each cluster, which is typically the dominant one, as well as the number of elements that form each cluster. Apart from Cluster 9, it is clearly noticeable that there is a medical speciality appointment 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.

The silhouette analysis performed is presented in Fig. 6. Keeping in mind that the silhouette score is an internal measure of cluster evaluation, measuring how well a sample is classified being assigned to a certain cluster, according to the inner-cluster tightness, as well as to the distance between different clusters [15], it is possible to verify that Cluster 9, having a negative average silhouette score across its samples, is composed by *outliers*. This conclusion is reachable due to the fact that the SSs of all samples in this cluster have very low values, being that the majority of the elements even have negative scores. 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 scores. 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,

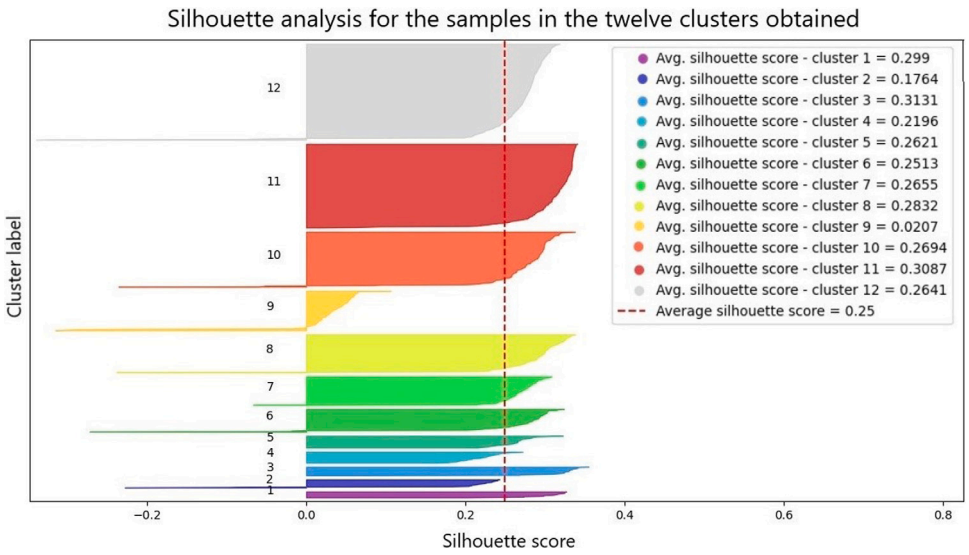


Fig. 6. Silhouette scores for each sample present in each of the obtained twelve clusters for $g = 0.1$ and $Tp = 10$, average silhouette score for each cluster and across all samples in the data set.

all remaining ones have an average SS higher than the overall score. Cluster 9 was expected to be below average since it is composed of *outliers*, while 2 and 4 were not. However, despite being below average, the difference is not much and may be justified by the smaller amount of elements that compose them.

So to obtain a more visual look at the clusters formed, a directed graph was developed for each cluster, where nodes represent medical speciality appointments 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. 7 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 appointments. Beyond the strong transition-wise relation between the first appointment from each cluster with Neurology appointments, it is possible to see that:

- Patients who start their activity in Orthopedics, apart from Neurology, get sent to Ophthalmology, Rehabilitation, and Anesthesiology appointments, more often than to other specialities present in Cluster 3 and quite a reasonable amount of patients transition from Neurology to Rehabilitation and Anesthesiology. It is also possible to detect a stronger transition-wise relation between Rehabilitation and GFM and subsequently to Orthopedics;
- The majority of patients who start off at Ob-Gyn appointments, apart from Neurology, more recurrently carry on to Surgery, Anesthesiology, GFM and Orthopedics appointments, compared to others. Not as often but still more recurrent than others, patients that form Cluster 5 transition from Neurology to ORL, Pneumology, Internal Medicine and Dermatology and from GFM and Pneumology to Neurology;
- Regarding Cluster 11, it is possible to identify transitions which stand out much more than others. For instance, it is common to find patients being redirected from Neurology to the top eleven most prevalent specialities after Neurology, being that the most prevailing ones are to Internal Medicine and Cardiology. Regarding incoming Neurology patients, these more often are being redirected from Internal Medicine, Cardiology, Ophthalmology and GFM. Transitions concerning other medical specialities besides the dominant one do not stand out in this subset;
- Finally, Cluster 12, the one with the most patients, is seen as a heterogeneous one, transition-wise. There are merely two transitions that stand out the most, which are transitions between GFM and Neurology appointments. A slightly more significant amount of patients in this cluster can also be seen transitioning from GFM to Cardiology, Orthopedics, Internal Medicine and Anesthesiology since these edges slightly stand out in the graph. Being the biggest sub-group of patients obtained, composed of 261 of them, it is natural to exist such diversity of transitions, remaining few of them that stand out. In addition, GFM is a heterogeneous medical speciality, justifying the fact that more specific patterns are not encountered.

Although the full clinical interpretation is still ongoing, we think that the discussion and interpretation of the obtained patient stratification already provide promising results regarding the usefulness of the method.

3.2.3. Variable screening and characterization per cluster

In order to attempt the identification of certain patient characteristics and patterns within the different clusters obtained, a variable screening, similar to the one completed initially, was put into practice.

Regarding the gender distribution per cluster, female predominance was observed in eleven out of twelve of the obtained subgroups. So to verify the significance of the gender proportions per cluster relative to the overall data set, the p-value from Fisher's exact test was calculated for each one of them. This allowed us to conclude that the Ob-Gyn and Anesthesiology clusters were the only ones with a significant gender proportion, presenting a p-value of 4.42e−8 and 0.01 respectively. The first one is formed entirely by female patients, while the second one has a slight male dominance.

Furthermore, distributions for patients' age and the number of chronic diseases were collected. Results are presented in Figs. 8 and 9, respectively. Regarding the age distributions, the Ob-Gyn cluster stands out from the overall age distribution initially obtained and from the remaining subgroups for being formed by patients with a lower age spectrum. The Orthopedics and Pneumology clusters are the ones that contain patients with a smaller variance in age (23 and 27 years, respectively) while the GFM one has the highest age variance of 45 years.

Moving on to the number of chronic disease distribution, the initial distribution presented in Section 4.1 indicated that the majority of Dementia patients suffered between two and thirteen chronic illnesses, from which 50% had at most four. This pattern is only observed in the Orthopedics cluster, whereas the remaining ones are composed of 50% of patients who have at most five or six co-existing chronic diseases. The Pneumology subgroup is formed by patients who suffer

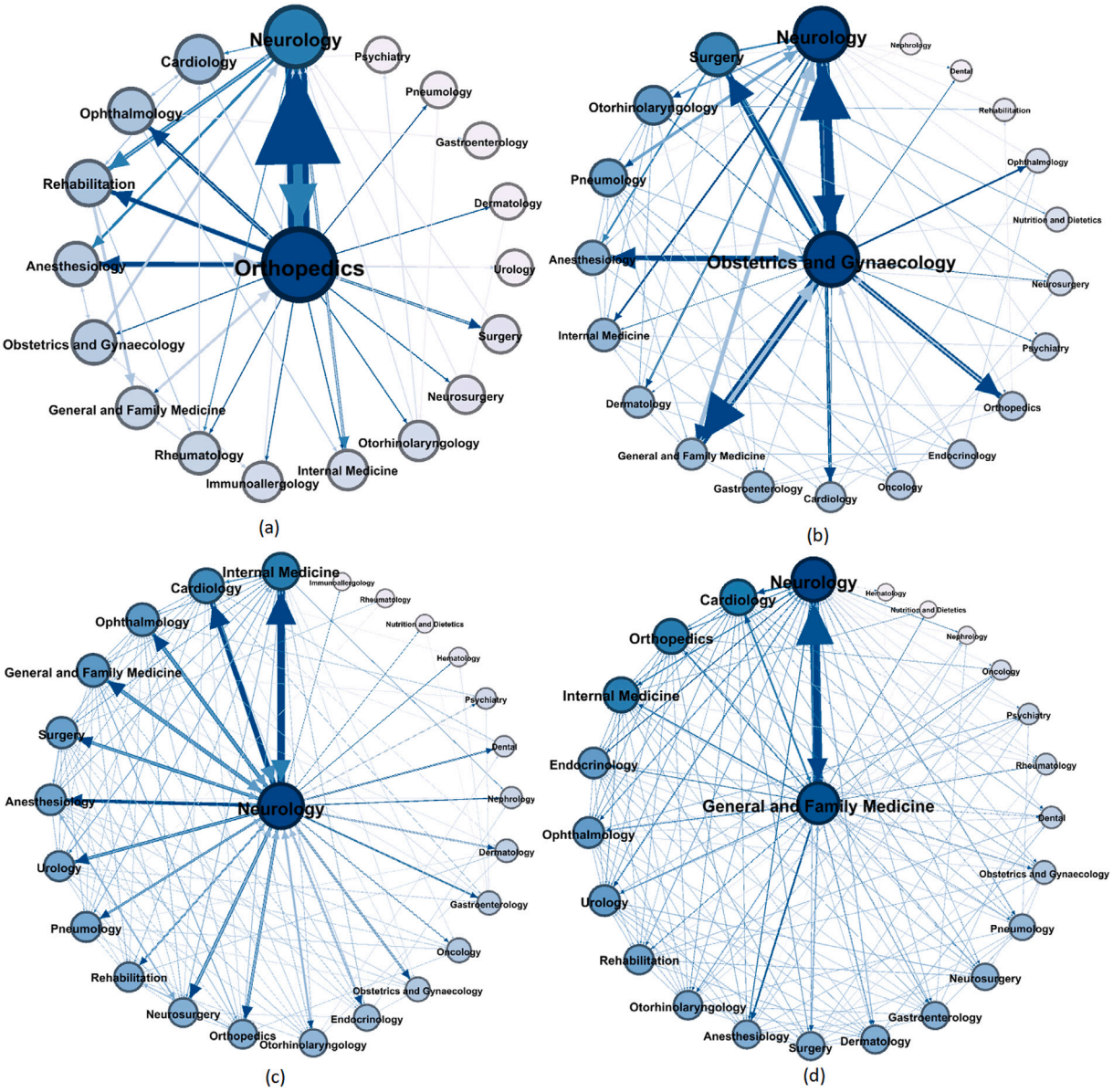


Fig. 7. Directed graphs showing the transitions between different medical appointments within clusters (a) 3, (b) 5, (c) 11 and (d) 12, where Orthopedics, Ob-Gyn, Neurology and GFM appointments prevail, respectively. Bigger and darker coloured nodes indicate a higher occurrence of that appointment in 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. As an example, consider the cluster (a), where we can see that patients who initiate their activity in Orthopedics are more likely to be referred to Ophthalmology, Rehabilitation, and Anesthesiology appointments compared to other specialities within the same cluster. Additionally, a significant number of patients transition from Neurology to Rehabilitation and Anesthesiology. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

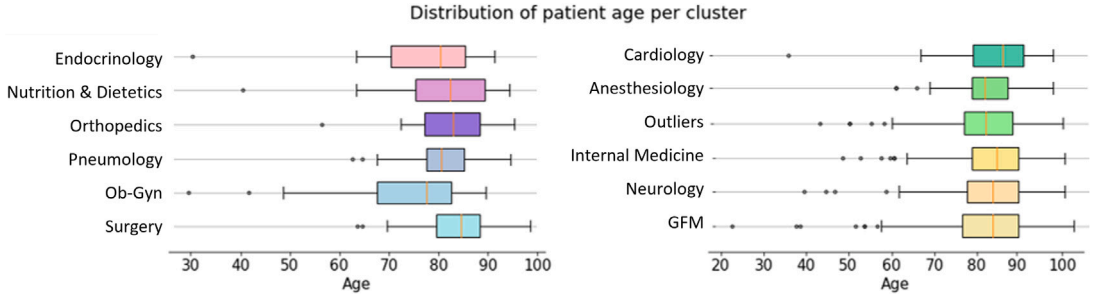


Fig. 8. Age distribution per cluster.

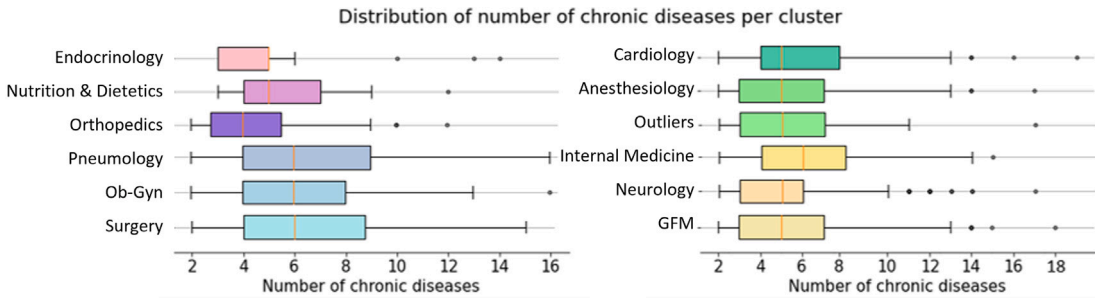


Fig. 9. Number of chronic diseases distribution per cluster.

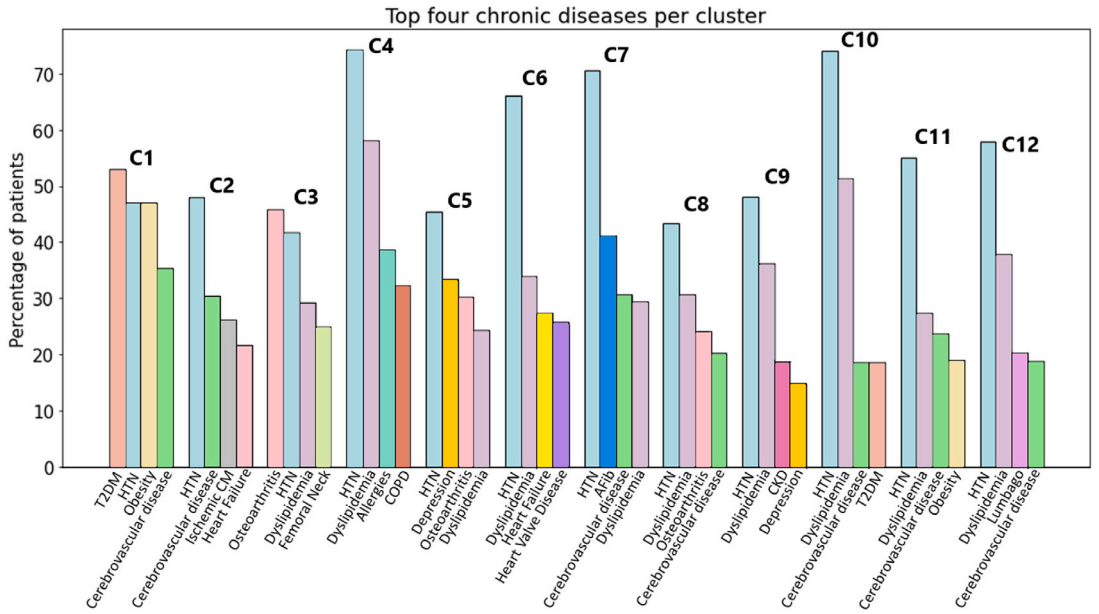


Fig. 10. Incidence of the four most prevalent chronic diseases per cluster, excluding Dementia. Each set of four bars represents the incidence in each one of the sub-groups, which are sorted, from left to right, in ascending order. The same disease is labelled with the same colour. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

from a wider range of illnesses, which may have up to sixteen of them. On the other hand, the Endocrinology cluster has the smallest range, since these patients mostly suffer between three and six co-occurring illnesses.

Furthermore, in order to understand the incidence of chronic diseases per partition obtained, the top four prevailing chronic diseases for each one of them were retrieved, which results can be seen in Fig. 10. With this analysis, we were able to identify some specific chronic disease patterns that can be associated with the prevalence of certain medical specialities in each subgroup. For instance, the most common chronic disease within the endocrinology cluster (C1) is type 2 diabetes, while within orthopedics patients (C3), the prevailing chronic illness is osteoarthritis. With respect to patients that belong to the pneumology cluster (C4), these have a high prevalence of allergies, and chronic obstructive pulmonary disease (COPD), which goes accordingly to their regular presence in pneumology appointments, since allergies are very often associated with respiratory issues and COPD is a lung-related disease.

Moreover, looking at the four most prevalent chronic diseases associated with regular attendance to surgery appointments (C6), besides HTN and dyslipidemia, these show a higher tendency of suffering from heart failure and heart valve disease, which may indicate that these diseases are more often related to surgical episodes when it comes to Dementia patients.

3.2.4. Hospital admission and emergency analysis

Data on 2078 hospital admissions (HA) and 9991 emergency episodes (EE) was available for this part of the study. Patients that did not belong to any of the obtained clusters were mapped out of the dataset, having pursued this analysis only with the hospital admission and emergency data belonging to the 1118 patients that integrate the clustering process. First of all, the fraction of these Dementia patients that actually had at least one of these occurrences and the average number of occurrences per patient was verified. In order to have a comparison baseline, the same data was obtained for the multimorbidity population in general. Results of this preliminary analysis are presented in Table 3.

Only 517 of the 1118 considered Dementia patients (46.24%) were admitted to the hospital in the considered time frame, while a much higher number of 931 patients (83.27%) had an emergency episode. Regarding all patients with MM, 15.82% of them registered HA, indicating that Dementia patients had almost three times the tendency of being admitted when compared to the general population with MM. Considering EE, 65.40% of MM patients had emergency episodes, corresponding to 1.27 more emergency events when it comes to Dementia patients. Considering now the average number of occurrences per Dementia patient, approximately two HA were identified, while for emergency episodes, that value was close to nine. For the MM population in general, an average of 0.24 HA and 3.06 EE were registered per patient.

Table 3
Percentage of multimorbidity (MM) and Dementia patients with hospital admissions (HA) and emergency episodes (EE) and corresponding average number of occurrences per patient, taking into consideration patients with zero occurrences, 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

Table 4
Percentage of Dementia patients, per cluster, with at least one episode of hospital admissions (HA) and emergency episodes (EE).

Cluster	Percentage of patients with HA	Percentage of patients with EE
Endocrinology	41.18	82.35
Nutrition and dietetics	65.22	91.30
Orthopedics	33.33	83.33
Pneumology	41.94	90.32
Ob-Gyn	42.42	75.76
Surgery	62.90	88.71
Cardiology	55.13	85.90
Anesthesiology	53.85	81.73
Outliers	35.18	66.67
Internal Medicine	55.33	87.33
Neurology	46.25	78.85
GFM	36.78	89.66

We saw that the average number of HA for Dementia patients between January 2007 and August 2021 was 7.75 times higher than for MM patients, while the average EE was almost three times higher. This goes accordingly to what is encountered in the literature regarding emergencies and hospitalizations of Dementia patients. For instance, Shepherd et al. [16] reached the conclusion that people suffering from Dementia are more frequently hospitalized when compared to those without this disease. They found that the risk of being hospitalized is 1.42 times higher for Dementia patients when compared to non-Dementia patients and that hospitalization rates oscillate between 0.37 and 1.26 per patient, per year [16]. Reasons for this higher tendency for HA include the age factor, the high incidence of comorbidities and the low independence and functional ability of these patients.

In order to understand whether belonging to a certain cluster could indicate a higher or lower probability of being admitted to the hospital or suffering an emergency episode, the percentage of Dementia patients within each cluster that had one of these events was assessed. Results can be observed in Table 4.

In a general way, we see that the probability of a Dementia patient having an emergency episode are very high, indicating that in one way or another, Dementia patients resort to emergency services quite often. Numbers regarding hospital admissions are not as high, however, they are also quite elevated, pointing to the conclusion that, roughly, fifty percent of this Dementia cohort have been at least once admitted to the hospital.

One of the highest percentages of HA is for Surgery patients, which is expected since, usually, when a patient undergoes a surgical procedure, they are admitted to the hospital. It is curious to observe the fact that the Nutrition and Dietetics cluster is the one formed by patients with higher percentages of both HA and EE, while Orthopedics patients register the lowest probabilities of HA events. A poor nutritional status of these patients may justify these high numbers for the Nutrition and Dietetics cluster. Regarding EE, the Ob-Gyn cluster registers the lowest fraction. Pneumology patients also present quite an elevated percentage of registered emergency episodes, as well as GFM, Internal Medicine, Surgery and Cardiology patients.

4. 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 features evaluation on the chosen cohort, both prior to the clinical pathway analysis, to better understand the patients under study, as well as subsequently to obtain 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 the most visited medical speciality appointments, as well as the most common transitions.

We were able to expose heterogeneity amongst patients, as well as activity patterns. The methods used in this work allowed us to pinpoint the prevailing attended appointments within Dementia patients, as well as to identify patterns when considering appointment transitions. When coupling this analysis to a variable 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 patient's 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 healthcare providers, and growing knowledge on how to deal with heterogeneous sources of healthcare data opens doors to new possibilities and advantages for the healthcare sector.

The results obtained in this work are probably not generalizable to all dementia patients worldwide, as numerous variables and confounding factors exist. It is typically challenging to ascertain the validity of results for vastly different populations. Nevertheless, this study included patients from a major private hospital in Portugal, and it would be interesting to determine whether these findings generalize to other populations. Apart from that, this study has some other limitations, for instance, the fact that only in-hospital data was considered for this study. Despite having a clinical pattern in hospital da Luz Lisboa, out hospital data was not accessible, meaning that some patient data might be missing from their medical pathway. Another limitation is the fact that the identification of multimorbidity patients might have a certain bias involved. Due to the fact that the dates of diagnosis were not considered, a particular patient might have lived without multimorbidity for some years, and only during the study period, a second disease was diagnosed.

Finally, we believe this pipeline has re-usability potential since it is easily adaptable to different cohorts. AliClu was originally developed with the aim of clustering patients based on their pharmacological treatment, focusing also on a certain disease. We adapted these methods to our cohort and goals, which were analysing clinical pathways. Hence, the pipeline can be adapted to different population groups, as well as to distinct events.

Future work which we believe would improve the performance of the methods used includes the use of natural language processing algorithms to extract information from clinical notes. Another interesting direction would be to adjust for specific variables, such as the age, besides exploring the relative weights of the pathways vs. the time between events. Additionally, to further validate the patterns identified among Dementia patients, it would be advantageous to test these methods in other disease datasets.

CRediT authorship contribution statement

Luísa Marote Costa: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **João Colaço:** Validation, Resources, Writing – review & editing. **Alexandra M. Carvalho:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Susana Vinga:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Andreia Sofia Teixeira:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge Jaime Machado and Carlos Magalhães, from Hospital da Luz, for the data extraction. This work was by the IntelligentCare project LISBOA-01-0247-FEDER-045948 that is co-financed by the ERDF/LISBOA2020, Portugal and by FCT, Portugal under CMU Portugal and FCT, Portugal through the INESC-ID and LASIGE Research Units, ref. UIDB/00408/2020 and ref. UIDP/00408/2020, UIDB/50008/2020, UIDB/50021/2020, DSAIPA/DS/0026/2019, PTDC/CTM-REF/2679/2020, PTDC/CCI-BIO/4180/2020. This project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 951970 (OLISSIPO project).

References

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

[2] J. Ryan, P. Fransquet, J. Wrigglesworth, P. Lacaze, Phenotypic heterogeneity in dementia: A challenge for epidemiology and biomarker studies, Front Public Health 6 (2018) http://dx.doi.org/10.3389/fpubh.2018.00181.

[3] World Health Organization, Dementia, 2021, https://www.who.int/news-room/fact-sheets/detail/dementia. (Accessed 28 September 2021).

[4] F. Bunn, A.-M. Burn, C. Goodman, G. Rait, S. Norton, L. Robinson, J. Schoeman, C. Brayne, Comorbidity and dementia: a scoping review of the literature, BMC Med. 12 (2014) http://dx.doi.org/10.1186/s12916-014-0192-4.

[5] B. Poblador-Plou, A.F. Calderón-Larrañaga, J. Marta-Moreno, J. Hancco-Saavedra, A. Sicras-Mainar, M. Soljak, A. Prados-Torres, Comorbidity of dementia: a cross-sectional study of primary care older patients, BMC Psychiatry 14 (2014) http://dx.doi.org/10.1186/1471-244X-14-84.

[6] K. Rama, H. Canhão, A.M. Carvalho, S. Vinga, AliClu - temporal sequence alignment for clustering longitudinal clinical data, BMC Med. Inform. Decis. Mak. 19 (2019).

[7] O. Ben-Assuli, R. Padman, I. Shabtai, Exploring trajectories of emergency department visits using a laboratory-based indicator of serious illness, Health Inform. J. 26 (2019) 205–217.

[8] H. Elghazel, V. Deslandres, K. Kallel, A. Dussauchoy, 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, http://dx.doi.org/10.1109/ICDIM.2007.4444236.

[9] R.C. Sato, D. Moraes Zouain, Markov models in health care, Einstein (Sao Paulo) (2010).

[10] S. Windgassen, R. Moss-Morris, K. Goldsmith, T. Chalder, The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome, J. Ment. Health 27 (2) (2018) 94–96, http://dx.doi.org/10.1080/09638237.2018.1437615, PMID: 29447026.

[11] A. Giannoula, A. Gutierrez-Sacristán, Álex Bravo, F. Sanz, L.I. Furlong, Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study, Sci. Rep. 8 (2018) http://dx.doi.org/10.1038/s41598-018-22578-1.

[12] A.Y. Wong, J. Karppinen, D. Samartzis, Low back pain in older adults: Risk factors, management options and future directions, Scoliosis Spinal Disord. 12 (1) (2017) http://dx.doi.org/10.1186/s13013-017-0121-3.

[13] M. Papaleontiou, M.R. Haymart, Approach to and treatment of thyroid disorders in the elderly, Med. Clin. North Am. 96 (2) (2012) 297–310, http://dx.doi.org/10.1016/j.mcna.2012.01.013.

[14] S.E. Jorgensen, B. Fath, Encyclopedia of Ecology, Elsevier Science, 2008, URL https://www.123library.org.

[15] S. Zhao, J. Sun, K. Shimizu, K. Kadota, Silhouette scores for arbitrary defined groups in gene expression data and insights into differential expression results, Biol. Proced. Online 20 (2018) http://dx.doi.org/10.1186/s12575-018-0067-8.

[16] H. Shepherd, G. Livingston, J. Chan, A. Sommerlad, Hospitalisation rates and predictors in people with dementia: a systematic review and meta-analysis, BMC Med. 17 (1) (2019) http://dx.doi.org/10.1186/s12916-019-1369-7.