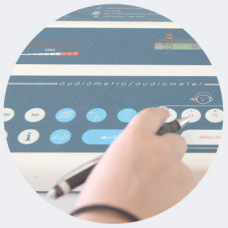
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA BIOMÉDICA






Hospital long-term care discharge clusters: a nationwide study using Clustering and Decision Tree methods

ANA RITA AFONSO CARREIRA
Novembro de 2022

P.PORTO

Hospital long-term care discharge clusters: a nationwide study using Clustering and Decision Tree methods

Ana Rita Afonso Carreira

Biomedical Engineer from Instituto Superior de Engenharia do Porto

"Dissertation presented at the Instituto Superior de Engenharia do Porto to obtain a Master's degree in Biomedical Engineering"

Advisers: João Vasco Santos, MD PhD

Goreti Marreiros, PhD

Co-Advisor: Alberto Freitas, PhD

November 2022

"The future belongs to those who believe in the beauty of their dreams."

Eleanor Roosevelt

Acknowledgements

À minha mãe e à minha avó, a quem devo tudo o que sou hoje. Obrigada por todo o vosso apoio incondicional e por acreditarem sempre em mim.

Aos meus orientadores, Professor João Vasco Santos, Professora Goreti Marreiros e Professor Alberto Freitas, por toda a orientação, por todos os conselhos preciosos e todo o tempo e empenho que disponibilizaram no decorrer deste trabalho.

Ao Professor Júlio Souza, ao Professor Diogo Martinho e ao Professor Vítor Crista por toda a ajuda e paciência, por todos os conhecimentos partilhados e por todo o tempo despendido no desenvolvimento deste projeto.

À Dr. Filipa Santos Martins por toda a ajuda, disponibilidade e por todas as sugestões.

A todos os meus amigos e colegas que sempre me apoiaram e me deram força e motivação para continuar este percurso.

Ao Instituto Superior de Engenharia do Porto por toda a formação ao longo destes anos, bem como a todos os professores que contribuíram para o meu crescimento pessoal e profissional.

Ao Centro de Investigação em Tecnologias e Serviços de Saúde (CINTESIS).

A todos os que direta ou indiretamente ajudaram a tornar este trabalho possível, o meu muito obrigada.

Abstract

Introduction: The ageing of the population structure leads to higher needs of long-term care (LTC). In order to adapt LTC and its associated policies it is important to establish the appropriate setting of personalised care. Hence, it is important to understand the associated factors that lead patients to the LTC use. The objective of this study is to assess clusters of hospitalised patients with higher proportion of discharges to LTC (LTCD) in Portugal, as well as to test the clustering method as a solution for an early identification of potential users, using different approaches.

Methods: A nationwide Portuguese study was performed, using inpatient data from Portuguese hospitals with discharges between 2012 and 2017. The variables used in this study were age, sex, principal diagnosis, comorbidities (identified using secondary diagnoses), admission type and hospital transfer. The main outcome of this analysis is being discharged to long-term and maintenance units (*Unidades de Longa Duração e Manutenção* - ULDM). Different approaches were applied to categorise principal diagnosis for each inpatient episode, using ICD-9-CM and ICD-10-CM main groups, ICD-9-CM and ICD-10-CM more detailed categories, Clinical Classification Software (CCS) and CCS Refined (CCSR). Subsequently, hierarchical clustering techniques were applied to determine the number of clusters in each dataset and decision tree methods were used to characterize each cluster.

Results: A total of 4427 inpatient episodes (0.23%) were discharged to LTC. Across the different methods to characterise principal diagnosis, the clusters with the highest proportion of discharges to LTC ranged between 0.7% and 60.8%.

Conclusion: There is great variability of the clustering results when comparing the different approaches of categorising principal diagnosis. The "quality" of the principal diagnosis categorisation overcomes the "quantity" (i.e. number of categories). This can have important implications for health system policies and hospital management Nevertheless, clustering methods showed to be good options to identify high-risk groups.

Key-words: long-term care; associated factors; hospital discharge; hospitalization; clustering; decision tree; Portugal.

Resumo

Introdução: O envelhecimento da estrutura populacional conduz a maiores necessidades de cuidados de longa duração. Para adaptar estes cuidados e as políticas associadas, é importante estabelecer o cenário adequado de cuidados para cada pessoa. Para isso, importa compreender os fatores associados à admissão de pacientes para estas unidades. O objetivo deste estudo é avaliar *clusters* de pacientes internados com maior proporção de alta para unidades de cuidados de longa duração em Portugal, bem como testar o método de *clustering* como solução para a identificação precoce de potenciais utentes, utilizando diferentes abordagens.

Métodos: Foi realizado um estudo nacional, utilizando dados de internamento de hospitais portugueses com altas entre 2012 e 2017. As variáveis utilizadas neste estudo foram idade, sexo, diagnóstico principal, comorbilidades (identificadas através de diagnósticos secundários), tipo de admissão e transferência hospitalar. Nesta análise, o principal resultado foi receber alta para Unidades de Longa Duração e Manutenção (ULDM). Diferentes abordagens foram aplicadas para categorizar o diagnóstico principal para cada episódio de internamento, usando os grupos principais da ICD-9-CM e ICD-10-CM, categorias mais detalhadas da ICD-9-CM e ICD-10-CM, Clinical Classification Software (CCS) e CCSR Refined (CCSR). Posteriormente, técnicas de *clustering* hierárquico foram aplicadas para determinar o número de *clusters* em cada conjunto de dados e métodos de árvore de decisão foram utilizados para caracterizar cada *cluster*.

Resultados: Um total de 4427 episódios de internamentos (0,23%) teve alta para ULDM. Entre os métodos, os *clusters* com maior proporção de altas para ULDM variaram entre 0,7% e 60,8% usando as diferentes categorizações de diagnósticos principais.

Conclusão: Há uma grande variabilidade dos resultados de clustering ao comparar as diferentes abordagens de categorização dos diagnósticos principais. A "qualidade" da categorização do diagnóstico principal supera a "quantidade" (ou seja, número de categorias). Isso pode ter implicações importantes para as políticas do sistema de saúde e gestão hospitalar. No entanto, os métodos de *clustering* mostraram-se boas opções para identificar grupos de alto risco.

Palavras-chave: cuidados de longa duração; alta hospitalar; hospitalização; clustering; árvores de decisão; Portugal.

Table of contents

List of figures

List of tables

List of abbreviations

ACS	Health Center Groups (Agrupamentos de Centros de Saúde)
ACSS	Central Administration of the Health System (Administração Central do Sistema de Saúde)
AHRQ	Agency for Healthcare Research and Quality
CCI	Charlson Comorbidity Index
CCS	Clinical Classification Software
CCSR	Clinical Classification Software Refined
CHF	Congestive Heart Failure
DM	Diabetes Mellitus without complications
DMcx	Diabetes Mellitus with complications
DRGs	Diagnosis Related Groups
DSP	Destination After Discharge
ECCI	Integrated Long-Term Care Teams (Equipas de Cuidados Continuados Integrados)
ECI	Elixhauser Comorbitidy Index
ECL	Local Coordination Teams (Equipas de Coordenação Local)
ECR	Regional Coordination Teams (Equipas de Coordenação Regional)
EGA	Discharge Management Team (Equipa de Gestão de Altas)
HCUP	Healthcare Cost and Utilization Project
HTN	Hypertension
ICD-10-CM/PCS	International Classification of Diseases - 10th Revision – Clinical Modification/Procedure Coding System
ICD-9-CM	International Classification of Diseases - 9th Revision - Clinical Modification

LTC Long Term Care

LTCD Long Term Care Discharge

PHTN Pulmonary Circulation Disorders

PUD Peptic Ulcer Disease

PVD Peripheral Vascular Disease

RNCCI National Network for Long-Term Care (Rede Nacional de Cuidados Continuados Integrados)

SNS National Health Service (Serviço Nacional de Saúde)

UC Convalescent Units (Unidades de Convalescença)

UCP Palliative Care Unit (Unidade de Cuidados Paliativos)

UCSP Personalized Health Care Units (Unidades de Cuidados de Saúde Personalizados)

ULDM Long-Term Maintenance Units (Unidades de Longa Duração e Manutenção)

UMDR Medium Term and Rehabilitation Units (Unidades de Média Duração e Reabilitação)

USF Family Health Units (Unidades de Saúde Familiar)

CHAPTER 1 – INTRODUCTION

1. Introduction

1.1. Contextualization and objectives

The increase in average life expectancy is one of humanity's best achievements. However, demographic changes such as the increase in the elderly population, the increase in chronic diseases and the increase in morbidity and functional restrictions are a reality in many developed countries [1]. There is a time in people's lives when their functional and physical abilities start to weaken. Thus, to continue to have a life as full and active as possible, they need help, whether from family, friends or qualified professionals for this purpose [2].

With this sense, in 1974, the National Health Service (SNS) was created in Portugal. SNS aims to promote universal and tendentially free access to health and to benefit all citizens and residents regardless of their financial situation and which extends throughout the territory [3], [4].

Before this, the concept of "long-term care" (LTC) already existed, since it emerged in 1963. LTC can be defined as a series of services required by people with a reduced degree of functional, physical or cognitive abilities and who, therefore, are dependent on help to perform their day-to-day basic activities for a long period [2], [5]. Therefore, it often consists of a combination of basic medical nursing services (dressings, medication, health monitoring and pain management) and palliative care, prevention, rehabilitation and family and social reintegration (Colombo et al., 2011; Decree Law nº 101/2006, 2006). Long-term care involves efforts to prevent the deterioration of the functional capabilities of disabled patients (such as the prevention of pressure ulcers and depression) by promoting appropriate lifestyles and adapting preventive care to maintain functional capabilities and social interaction. The population in need of long-term care services includes all patients suffering from any type of physical, psychological and/or mental disability [7].

Thereby, the objective is to ensure that the individual can maintain the best quality of life, with the highest degree of independence, autonomy, participation, and human dignity. The appropriate long-term care includes respect for individual values, preferences and needs, and may be in their own home or an institution. The type of care

needed, and its duration are the most difficult to predict. Many people may actually regain their functional abilities, through rehabilitation [8].

With these objectives, in Portugal, the National Network for Long-Term Care (Rede Nacional de Cuidados Continuados Integrados - RNCCI) was created in 2006 as a partnership between the Ministry of Health and the Ministry of Labour and Social Solidarity. Long term-care can be provided by several units: inpatient units, outpatient units, hospital teams and home care teams. The inpatient institutions are divided into three units, depending on the type of care needed and the estimated length of stay: Convalescent Units ("Unidades de Convalescença" – UC) for admissions with the predictability of up to 30 days; Medium Term and Rehabilitation Units ("Unidades de Média Duração e Reabilitação" – UMDR) for inpatients between 30 to 90 days; Long-Term Maintenance Units ("Unidades de Longa Duração e Manutenção" – ULDM) for a period of hospitalization longer than 90 days (Decree Law nº 101/2006, 2006).

Regarding the Long-Term Maintenance Units (Unidades de Longa Duração e Manutenção – ULDM), these consist of inpatient units, temporary or permanent, with the aim of providing social support and maintenance health care to people with chronic illnesses or processes, with different levels of dependence and who do not have conditions to be cared for at home. These units have the intent to provide care that prevents and delays the worsening of the dependency situation and ensure, namely maintenance and stimulation activities, daily medical and nursing care, prescription and administration of drugs, psychosocial support, periodic physical control, care physiotherapy and occupational therapy, sociocultural entertainment, and support in activities of daily living [6].

Concerning the referral process, the individuals can be referred to the RNCCI through hospitals or through primary health care. In 2019, according to RNCCI monitoring reports, more than 80% of referrals were from hospital discharge teams [9]. If they are proposed through hospitals, the identification and signalling is handled by the Discharge Management Team (Equipa de Gestão de Altas - EGA) which prepares and manages hospital discharge and ensures the referral of the patient's process, taking into account their needs, degree of dependence, possibility of rehabilitation and their socio-family context. The request is sent to the Local Coordination Teams (Equipas de Coordenação Local – ECL) that evaluates the individual health and social situation, checking whether the patient meets the criteria for RNCCI.

On the other hand, if they are proposed through primary health care, the request is sent to the ECL that is concerned with evaluating and validating the patient's referral. If the patient meet the criteria and the referral is approved by the ECL, the request is directed to the Regional Coordination Teams (Equipas de Coordenação Regional - ECR) that is responsible for referring patients to the appropriate level of care at the RNCCI [1], [6], [10].

Hospital admission of a patient with LTC needs may occur due to an acute illness, complication or evolution of a chronic disease or even deterioration of the health status that requires medical and social management [11]. Hospital discharge should be thus planned as soon as possible, to ensure continuity of care and efficient use of hospital and community resources [12]. Also, hospitals are under great pressure to shorten lengths of stay [13] while patients with LTC needs usually present longer lengths of stay [14], [15].

The information on all inpatient episodes in all public hospitals in mainland Portugal, as anonymous information, is centralized by the Central Administration of the Health System (Administração Central do Sistema de Saúde – ACSS) using the Hospital Morbidity Database. For each discharge there is information regarding the birth date, sex, district and municipality of residence, but also date and type of admission, the date of discharge, the days of hospitalization, hospital identification, admission outcome (death, discharge or transfer), Diagnosis Related Groups (DRGs) and codes for primary diagnosis, secondary diagnosis, procedures, and external causes [16], [17].

This clinical coding consists of a process of categorizing information contained in clinical records, referring to outpatient episodes, outpatient consultations, long-term hospitalization or emergencies, through alphanumeric codes related to diagnosis, procedures, external causes and other information about the patient [18]. In Portugal, this process is carried out by medical doctors with specific training to guarantee uniformity of this process and ensure that its registration is carried out in a similar way by all coders in all hospitals in the country [17].

After entering the data from the discharge clinical record, the coders assign a clinical code according to coding systems, such as the International Classification of Diseases - 9th Revision - Clinical Modification (ICD-9-CM) or International Classification of Diseases - 10th Revision – Clinical Modification/Procedure Coding System (ICD-10-CM/PCS). In Portugal, ICD-9-CM was used between 1989 and 2016 and ICD-10-CM/PCS began to be used in October 2016 in some pilot hospitals, such as

Centro Hospitalar Lisboa Central, Centro Hospitalar de São João and Hospital do Espírito Santo – Évora. On January 1, 2017 the ICD-10-CM/PCS went into effect in all hospitals in the country [19], [20].

The coding system ICD-9-CM encompasses a set of diagnoses codes and procedures codes used for classifying and coding information on hospital morbidity, for mainly reimbursement purposes through DRGs [21]. The transition from this system to ICD-10-CM/PCS allowed consistency between assigned codes and advances in medical technology, facilitating international comparisons of healthcare quality, allowing to share global best practices and helping to increase payment accuracy. This system with an increase of codes and the number of characters of each code brought more specificity, accuracy, precision and veracity to the generation of higher quality data, contributing to a more accurate measurement of the quality, safety and effectiveness of health care and helping to detect public health diseases more effectively, more completely and clearly than the ICD-9-CM. Thus, these data not only are essential for carrying out epidemiological and health services research but also to allow the evaluation of hospital production and quality [17], [20].

In 1999, in order to help clinical research using ICD-9-CM codes, the Agency for Healthcare Research and Quality (AHRQ) introduced the Clinical Classification Software for ICD-9-CM (CSS). This software allows to group and categorize ICD-9-CM codes, which are over than 14 000 diagnosis codes and 3 900 procedure codes, into a smaller number of categories. There are CCS single level with 285 diagnosis categories and 231 procedure categories and multi-level CCS with 4 levels in the multi-level diagnosis CCS and three levels in the multi-level procedure CCS [22]. On the other hand, for data using ICD-10-CM/PCS codes, which gathers more than 70 000 diagnosis codes and 80 000 procedure codes, the AHRQ developed the Clinical Classification Software Refined (CCSR) with over 530 diagnosis categories and over 320 procedure categories [23]. These approaches are useful to present descriptive statistics analysis for research, and can also be applied for risk factor identification, risk-adjustment, or for benchmarking hospital quality (e.g. inpatient mortality) [24]–[26].

In this thesis, the aim is to assess clusters of hospitalizations with higher probability of LTC discharge (LTCD) and their characteristics. This will be done by using clustering and decision tree methods, which were applied to data from mainland public hospitals in Portugal.

1.2. Dissertation structure

This dissertation is divided into six chapters. This first chapter introduces the context and objectives of the work. In addition, the structure of the dissertation is also defined here.

Chapter 2 is the literature review of the main themes of this work. Some literature studies and analyses conducted in the past with similar objectives or using similar methodologies are described.

Subsequently, in chapter 3 the methodology used in this study is explained. This section includes the definition of the sample, including the inclusion criteria, as well as the considered dataset and variables, the data pre-processing and the data analysis.

The chapter 4 presents the results obtained. The results presented are relative to the descriptive data analysis and the cluster and decision tree analysis.

The chapter 5 refers to the discussion and analysis of the results.

Finally, in the chapter 6 there are presented some conclusions of this work as well as limitations and some suggestions for future studies.

CHAPTER 2 – LITERATURE REVIEW

2. Literature Review

2.1. Long-term care determinants

In Europe, the amount of people in need of long-term care will increase dramatically in 2060. This proves to be an urgent public health problem and the health care system needs to be prepared to solve it and redefine new policies to establish the appropriate setting of care for each person. For this, it is important to understand the main risk factors for being placed and discharged for long-term care over time. Through the literature, it is widely mentioned that those factors include individual characteristics such as sociodemographic, medical conditions and physical and cognitive dependence levels [1], [27].

A study developed by Kuzuya et al., allowed to identify long-term care predictors through data of 1 739 dependent older adults after a follow-up period of 36 months from an institution in Nagoya, Japan. Data included demographic characteristics, basic activities of daily living, comorbidities, and use of home care services. They performed an analysis using Kaplan-Meier curves and multivariate Cox proportional hazards models. Results showed that the average age is 84,4 years and that 33,1% of the study population were men. It was also possible to understand that there was a high prevalence of cerebrovascular disease (42,8%) and dementia (44,2%) and that 17,3% lived alone [28].

Wu and his colleagues carried out a population-based study in Taiwan in order to identify the determinants of long-term care services among the elderly people and to differentiate the characteristics of people using home/community-based services and institution-based services. The population sample was 2 608 individuals with 65 years and over and the data was from interviews of the national health system of Taiwan in 2005. The users of institution-based services proved to be less educated, more likely to be single and have fewer family members, higher prevalence of stool incontinence, dementia, and higher level of disability. The authors were able to suggest advanced age, stroke, dementia, difficulty in the activities of daily living and being single as the main factors for LTC need. Regarding the influence of the sex, they did not find significant differences in age by sex, however women showed a higher level of disability, more geriatric conditions and higher prevalence of hypertension, diabetes, hyperlipidemia, and

dementia than men. This can be explained as women tend to live longer than men. Thus, it is important to consider the gender disparity in terms of health and social conditions when providing long-term care [29].

Through a systematic review and meta-analysis of observational studies from Europe, North America and East Asia, Burton et al. established predictive factors for discharge to institutional long-term care after stroke. Their results showed that older age and greater stroke severity were associated with being more likely to long-term care admission. In regard to the sex, there was no evidence of being a predictor to the need for long-term care. Furthermore, they explored the association with previous stroke, comorbidities, dementia and delirium or complications during inpatient stay and these had high rates of long term-care admission [30].

In Germany, a study was developed to identify the determinants for utilization and transitions to long-term care in adults aged 65 and more between 2011 and 2012 and 2016, from population-based Cooperative Health Research in the Region of Augsburg. The authors used generalized estimating equation logistic models to identify the determinants and types of long-term care services and a logistic regression model to analyse the determinants of transitions to continuing care over four years through a longitudinal analysis. The determinants analysed were related to predisposition (age, sex, education), qualification (housing, income) and need (multimorbidity, disability and inability). With a sample of 810 individuals, they determined that the predisposing factors were advanced age (mean age of 78,4 years), being female and having high multimorbidity and disability factors. In addition, determinants such as living along, high incomes and high level of disability proved to be significant for the choice of formal long-term care services [27].

Chen et al. analysed the characteristics and predictive factors influencing patients using long-term care services of discharge planning using Andersen Behavioral Model through a hospital-based cross-sectional study in Hualien, Taiwan during November 2017 and October 2018. Among the total of 280 patients, the results showed that the vital factors that influenced the use of the long-term care services were the age (average of 75,1 years), the medical accessibility (58,5% did not live near medical institutions), suffering from chronic obstructive diseases or asthma (67%), cerebrovascular disease or stroke (82%), coronary or cardiovascular (79,2%) and diabetes mellitus (75,8%). Among them, the factor that proved to be the most influential was the age. Female gender

represented 65,8% of the studied population. The authors did not find a significant association between gender, tubal insertion, dementia, osteoarthritis, malignancy or last stage of kidney disease and the use of long-term care [31].

Momose et al. performed a cross-sectional study with 2013 data from a Comprehensive Survey of Living Conditions conducted by the Ministry of Health, Labour and Welfare of Japan. Through univariate and multivariate logistic regression analyses, the authors were able to determine that the factors that showed to be more associated with long-term care included older age, the interaction between sex and age between 85 and 89 years, difficulties in limb movement, swollen and heavy feet, incontinence, severe psychological disorders, dementia, stroke, Parkinson's disease, chronic obstructive pulmonary disease, fractures, rheumatoid arthritis, kidney disease, diabetes and osteoporosis. However, the factors that had a negative influence were the presence of a spouse, regular visits to the hospital due to hypertension and turning to friends about worries or stress [32].

In Portugal, Coutinho carried out a qualitative and quantitative retrospective study of patients admitted to the services of the HUC (Hospitais da Universidade de Coimbra) of CHUC (Centro Hospitalar e Universitário de Coimbra), referred and admitted to RNCCI during 2016, as well as professionals from medical specialties and social services. The author collected data from referral forms for the RNCCI, social information on patients, from Hospital Management Base of the HUC and semi-structured interviews with health professionals. As for the results, in 2016, inpatients referred to RNCCI revealed to be similar in terms of sex, with 51% being male. In terms of age, mostly were in the 80-89 age group (36,8%). Regarding the marital status, 46,2% were married and 36,1% were widowed. Considering educational qualifications, 44,1% had completed the first cycle and 17,4% could not read or write. Respecting to the main causes of hospitalization, 19,3% were bone fractures, 18,4% stroke, 16,9% oncological disease and 7,4% respiratory infections, mainly occurring in Orthopedics Services, Internal Medicine A and Neurology. The most requested units to which the patients were referenced within the RNCC were UC (42,4%), UMDR (29,3%) and ULDM (13,4%). The least requested were Palliative Care Unit (Unidade de Cuidados Paliativos – UCP) with 10,7% and Integrated Long-Term Care Teams (Equipas de Cuidados Continuados Integrados – ECCI) with 4,1% [10].

In 2020, the same authors performed a study with the objective to identify the main risk factors related to the probability of the individual being placed in different units of RNCCI (UC, UMDR and ULDM) and how the individuals differ from each other. The authors used data from the Central Health System Administration and applied a logistic regression to identify the determinants of admission to long-term care services with demographic characteristics, medical conditions, and levels of dependence as independent variables and as control variables the region of care, the reference entity, and the placement process. In order to identify the contribution of these factors in each unit, an ordered logistic regression was used. Therefore, the results showed that the main predictors for the individual to be institutionalized were being female, not being married, having low social support, being literate, mental illness and diseases of the circulatory, nervous, or musculoskeletal system. On the other hand, old age and the existence of family or neighbour support have the opposite effect. Considering the level of dependency at the time of admission, being classified with the higher level of cognitive and physical dependency decreases the possibility of being admitted to an institution. It was also understood that dependency levels increase across units, from UC to ULDM. Regarding the ULDM, old age, suffering from cancer or mental illness were the main predictors to an individual being admitted in this unit [1].

Macedo analysed 2019 data from user processes referred to the RNCCI by the EGA of Medical Service of the Pombal District Hospital with the aim to analyse the role of long-term care services in the articulation between health and informal social support. From a total of 57 patients, 34 were admitted to Continuing Care Units and of these, 25 patients were discharged from units of the RNCCI. The average age were 80 years (varying between 50 and 93 years) and 63,2% were females. All patients lived in their own house, except 3 who lived on a house rented or lent by a family member. As for residence, it is known that 64,9% lived in rural areas. 56,1% of the individuals were married and 31,6% widowed, with only 7% single and 5,3% divorced. Of the 57 patients, 58% met criteria for admission to UMDR. 21% for UC, 12,2% for ULDM and 8,8% for ECCI. 61,8% of the patients who were discharged from the RNCCI were female and 38,2% male [33].

According to the Monitoring Report of the RNCCI carried out by the Management Department of the Health Services and Resources Network (Departamento de Gestão da Rede de Serviços e Recursos em Saúde – DRS) of the ACSS, the number of users

referenced in 2019 was 43 750. 90% of these were referred due to dependency on activities of daily living and 89,5% due to informal user/caregiver education. Overall, the main reasons were the need for psychosocial rehabilitation (73,4%) and the need for medication management (70,5%). Regarding the referencing process, 63,2% of individuals were referred by hospitals and 36,8% by Primary Health Care (CSP), with an increase compared to 2018.

Concerning the characterization of the users, 84,4% (83,7% in 2018) of the population aged over 65 years and in ECCI this value was 86,1% (85,4% in 2018). Of the total number of individuals, 51,1% were aged over 80 years and 53,4% were in ECCI. Regarding the sex, females represented 56,2% of all users, 50,4% of the users were female over 65 years old and 65,1% were female over 80 years. As for the level of education, 22,3% had no instruction and 66,1% had education between 1 and 6 years. About the marital status, 13,8% were single and 32,1% were widowed. In relation to the cohabitation, 69,7% lived with their natural family and 25,1% lived alone.

In relation to diagnosis, aggregating main and secondary diagnosis, 12,7% of the main diagnosis were related to acute but ill-defined cerebrovascular disease, 5,5% to cerebrovascular disease not classifiable elsewhere or ill-defined, and 1,5% to intracerebral haemorrhage. With respect to the diagnosis related to cerebral vascular pathology (adding to the above mentioned, the late effects of cerebrovascular disease, occlusion of cerebral arteries and unspecified intracranial haemorrhage), they accounted to 22.5% of the diagnosis. The fracture of the femur neck associated with the diagnosis fracture of not classifiable elsewhere parts or unspecified parts of the femur represented 11,8% of the diagnosis and the diagnosis of chronic skin ulcer represented 6,6% of the total of main and secondary diagnosis [9].

2.2. Cluster analysis, decision tree methods and Hospital Morbidity Database

Clustering was introduced as an unsupervised classification of patterns into groups to the data mining research. Therefore, clustering techniques are used to decompose a set of individuals into natural groups and find groups that are lightly connected with each other [34].

For instance, Brugnaro et al. developed a study to determine the prevalence and the risk factors for Methicillin-Resistant Staphylococcus Aureus (MRSA) carriage in

nursing home residents in Vicenza, Italy. The authors performed a point prevalence survey in two long-term care facilities from 12 June to 6 July 2006. The factors found to be significantly associated with MRSA carriage at univariate analysis were introduced into multilevel logistic regression models to estimate the odd ratios (OR) with 95% of confidence interval (CI) for the risk of MRSA colonization, considering clustering of patients within wards. The study group consisted of 551 patients, 73% of them were female. The mean age was 83 years and 31% were at least 90 years old. Overall, residents were highly dependent, having limitations on all activities of daily living and 71% of the patients were incontinent, but only 6,5% had a urinary catheter. The most common invasive device was a gastrostomy tube (10%). 21% of the residents had been admitted to hospital at least once in the preceding year [35].

With the aim to examine the Active-Ageing concept in the context of residential long-term care facilities and the determinants within this setting, van Malderen et al. carried out a qualitative study with semi-structured focus groups and a thematic analysis. The sample included four focus groups of 8 residents of long-term care facilities, 8 children of residents, 8 community-dwelling older people and 6 gerontologists. The data were deconstructed and reorganised by clustering. This allowed a cluster-process based on the expectations of the researcher (through literature and theories – deductive analysis) and on the revealing data (inductive analysis). This analysis was based on the determinants of the WHO-document on Active-Aging and when codes could not be assigned based on these determinants, but were related, new clusters were developed within the WHO-determinants. This comparison and clustering method allowed the analysis and understanding the data and generated nine determinants of AA. Seven determinants correspond to those identified by the WHO: Culture, Behaviour (Tobacco use, Physical activity, Healthy eating, Oral health, Alcohol, Own decision), Psychological Factors (Cognition, Self-efficacy, Coping, Finding peace), Physical Environment (Environment, Safe housing, Homeliness/personality, Normalisation), Social Environment (Social support, Violence and abuse, Education and literacy, Group living, Communication), Economic Characteristics (Social protection, Work) and Health and Social Care (Physical care, Psychological care, Individualised care, Coordinated care). Two new determinants were identified: Meaningful Leisure and Participation (Activating residents, Control of own life and care, Participation in organisation and functioning of the nursing home) [36].

Igarashi et al. developed a study with the aim to classify patterns of inpatient characteristics among Japanese long-term care wards and hospitals and to examine their functional characteristics. They used data from 1856 long-term care wards from the 2014 Annual Report for Functions of Medical Institutions in Japan. For this, the authors conducted a descriptive analysis to perform a case-mix classification of hospital/ward characteristics. Case-mix classification is usually used in payment systems to refund health care providers based on the type of patient, in risk adjustment models for health outcomes or in other quality measures, staffing, program evaluation and long-term planning and budgeting for policy makers [37]. Then, the authors did a cluster analysis based on the proportion of patients categorised for each of 9 case-mix classifications. From this analysis, resulted 5 clusters with the best conceptual fit and they compared the characteristics of these clusters with chi-square tests and analysis of variance. As for the results, cluster 1 was low medical acuity/high activities of daily living; cluster 2 was medium medical acuity/high activities of daily living; cluster 3 was medium medical acuity/low activities of daily living; cluster 4 was high medical acuity/low activities of daily living; and cluster 5 was mixed. With this study, the authors were able to verify that the inpatients within clusters with high activities of daily living needed support in promoting home discharge. On the other hand, clusters with low activities of daily living needed support in providing quality end of life care [38].

A study to assess the efficacy of interventions to prevent delirium in older people in long-term care institutional settings was performed by Woodhoyse et al. The authors included randomised controlled trials (RCTs) and cluster-randomised controlled trials (cluster-RCTs) of single and multicomponent, non-pharmacological and pharmacological interventions for preventing delirium in older people in permanent LTC residence. For this purpose, they used risk ratios (RR) with 95% confidence interval (CI) as measures of treatment effect for dichotomous outcomes and hazard ratios (HR) with 95% CI for time to event data. The effect measures and their 95% CIs that were adjusted for clustering were extracted for cluster-RCTs. If unadjusted analyses had been performed, they calculated approximately correct analyses, by extracting data on number of clusters, mean size of each cluster, primary outcome data and estimated the intracluster correlation coefficient (ICC). Moreover, if an approximately correct analysis was not possible, then the authors extracted primary data and calculated RRs with 95% CIs. Considering the results, three RCTs of delirium prevention interventions for older people in institutional

LTC were identified. One small RCT group consisting of 98 participants of a hydration-based intervention was unable to show any reduction in the incidence of delirium in the intervention group compared to the control due to a very serious inaccuracy in the outcome. Moreover, one large cluster-RCT with 3538 participants of a computerised system to identify medications that may contribute to delirium risk and trigger a pharmacist-led medication review found moderate-certainty evidence of a large reduction in delirium incidence but of little or no effect on hospital admissions, mortality or falls. In addition, one feasibility cluster-RCT of 215 participants of an enhanced educational package to identify delirium risk targets and develop bespoke solutions specific to individual care homes, was not able to show any reduction in delirium incidence or prevalence due to the serious imprecision in the results [39].

In order to understand the representations of elderly Europeans in their place of residence, Chrusciel et al. performed research in adults aged over 65 in four European countries. For this purpose, the authors carried out a cross-sectional study by a poll institute with a representative sample of individuals in four European countries, where a total of 4160 subjects were selected. Then, for descriptive analysis, qualitative variables were stated as absolute frequencies with proportions and the results of the questionnaire were analysed using Principal Components Analysis (PCA) and six clusters of residents were identified from this analysis. PCA is a multivariate statistical technique that allows to analyse a data table with observations with several dependent variables, generally inter-correlated. This technique permits to extract principal information and express it as a set of new variables (principal components) [40]. The authors were able to state that overall, 57% of the individuals were women, 34% were aged between 65 and 69, 65% were living maritally, 70% owned their home, 30% showed signs of social precariousness including low income and 27% had low level of education. The six defined groups were related to Wealthy Belgians, Flexible single people, Wealthy Germans, Low-income Germans, Isolated Italians, Italian homebirds, respectively [41].

On the other hand, Boockvar et al. conducted a study with the aim of testing the effectiveness of a Hospital Elder Life Program in Long Term Care (HELP-LTC) for long-term nursing home residents in a single-site cluster randomized controlled trial. Therefore, 219 nursing home residents who developed an acute illness or change of condition were randomly assigned to HELP-LTC (n = 105) or usual care (n = 114) per unit. The primary delirium outcome and delirium severity were checked each weekday

by a research assistant for assignment to a group, using the Confusion Assessment Method (CAM) and CAM severity score (CAM-S), respectively. Cognitive function was determined using the Cognitive Performance Scale (CPS). They carried out the verification of hospitalization through the review of medical records. The inclusion criteria for choosing the study participants were: 1) living in a long-term care unit; 2) are expected to survive for at least 2 months and have no plan to be discharged, and 3) suffer from an acute illness or change in condition. As for the results, the mean age was 81.7 years (standard error corrected by grouping (SE): 1.1) and 65.3% were female. The usual care group showed to have a higher frequency of heart failure than the intervention group (34.3% vs 23.7%; p = 0.011) but had a similar number of chronic diseases (2.8 vs 2, 9; p = 0.809). The most common acute diagnosis in both groups were urinary infection (17.8%), skin infection (8.7%) and respiratory infection (7.8%) [42].

About the methods used in this study, clustering techniques and decision tree methods were applied to assess clusters with higher proportion of LTC discharge (LTCD) and their characteristics. This hybrid model was previously used by Martinho et al. with the aim to identify and classify different groups of patients with chronic obstructive respiratory diseases. Combining hierarchical agglomerative and divisive clustering and classification models, the authors were able to describe 5 clusters [43].

Regarding the database, Scripcaru used data from the Hospital Morbidity Database (BD GDH) of the Central Administration of the Health System, I.P. (ACSS) from 2004 to 2013, with the objective to identify and characterize drug-related adverse events in a hospital context. The events were identified based on Diagnosis Related Groups (DRG – GDH in Portuguese). A global analysis model was developed, and all clusters showed increasing trends, for the period 2004 to 2013. Through a spatiotemporal clustering analysis, the author was able to detect where and when an unexpected frequency of a given health phenomenon was happening, as well as this analysis was the basis to more complex multivariate processes allowing the incorporation of factors of risk. For this retrospective spatiotemporal analysis, the Poisson distribution was used to identify the high risk of events within a cluster, using SaTScan software. The cluster analysis allowed them to detect and identify geographic areas that had significant differences in risk, regardless of their size. As for the SaTScan software, it is based on Kulldorff's statistical methodology, uses the spatial scanning technique and is commonly used in Public Health [17].

Similarly, Lacerda et al. also used data from the Hospital Morbidity Database in order to evaluate hospital inpatient care in the National Health Service in mainland Portugal by pediatric patients with complex chronic conditions, between 2011 and 2015. The authors conducted an observational longitudinal retrospective epidemiological study and were able to conclude that 15,5% of admissions contained at least one complex chronic conditions code and they represented 87,2% of deaths, 39,4% of expenses and 29,8% of hospital days [44].

With the objective to describe and model a time series of hospital admissions for diabetes in Portugal, Fialho also used data from the Hospital Morbidity Database. The author selected all diagnoses of diabetes as the primary cause of admission, coded to the third digit by 250 (diabetes mellitus), according to ICD-9-CM or E10 (type 1 diabetes), E11 (type 2 diabetes), E13 (another type of diabetes), according to the 10th ICD Revision (ICD-10-CM/PCS). These records were associated with a specific episode and the ones selected had an admission date between January 1, 2010 and December 31, 2018 and one day of minimum hospital stay. Thus, the author collected a total of 108 episodes and used a subset of 84 episodes to identify and estimate the model (training set) and the remaining episodes for the test set and model validation. The variables from Hospital Morbidity Database used for this study were sociodemographic (sex, age, region of residence) and also clinical variables (mode of admission, length of hospitalization and patient destination after discharge) [45].

CHAPTER 3 - METHODOLOGY

3. Methodology

The present work was developed with the aim is to assess clusters of hospitalised patients with higher percentage of discharges to LTC (LTCD) in Portugal, as well as to test the clustering method as a solution for an early identification of potential users, using different approaches. To achieve this, it was necessary to resort to data from several mainland public hospitals from Portugal that report an expected number of discharges to ULDM and to use clustering techniques.

Initially, it was necessary to define the sample of population to be used in the study and to analyse and describe the data based in the characteristics of the population. Then, the data needed to be pre-processed to be able to be applied in the clustering method used to determine the clusters.

3.1. Study design and data sources

A retrospective observational study was conducted using a national hospitalisation database, namely the Portuguese Hospital Morbidity Database, which included coded clinical data from all inpatient episodes provided by all mainland Portuguese public hospitals, discharged between January 1st 2012 and December 31st 2017. This administrative database was provided by the Central Authority for Health Services (Administração Central do Sistema de Saúde - ACSS) of the Portuguese Ministry of Health. Each hospitalisation was considered as an independent episode.

Considering that the data had some episodes coded in ICD-9-CM and others in ICD-10-CM the data was divided in two datasets, to facilitate the analysis: one with episodes between 2012 and 2016, only with data coded with ICD-9-CM; other with data from 2017 coded with ICD-10-CM. In each, only data of inpatient episodes, statistically valid, from the selected hospitals, which had an LTC service as discharge destination (DSP=63) were considered.

3.1.1. Inclusion Criteria for Hospitals

First it was necessary to calculate the proportion of hospitalisations that had as destination after discharge a LTC service for each hospital. Considering that it is

important to choose hospitals with a considerable proportion of discharge to LTC and a good data quality, these proportions and the relative difference between the number of users, referred to the RNCCI Database and the Hospital Morbidity Database, were used to determine the hospitals to be included in this study. The results obtained with this are represented in Figure 1. Thus, to be included in this study, it was established that hospitals should meet the following quality criteria: relative difference of referencing to RNCCI and Hospital Morbidity Database between -10 and 10. Therefore, the hospitals represented with the numbers 14, 15 and 21 were excluded and the number of hospitals used in this analysis was 19.

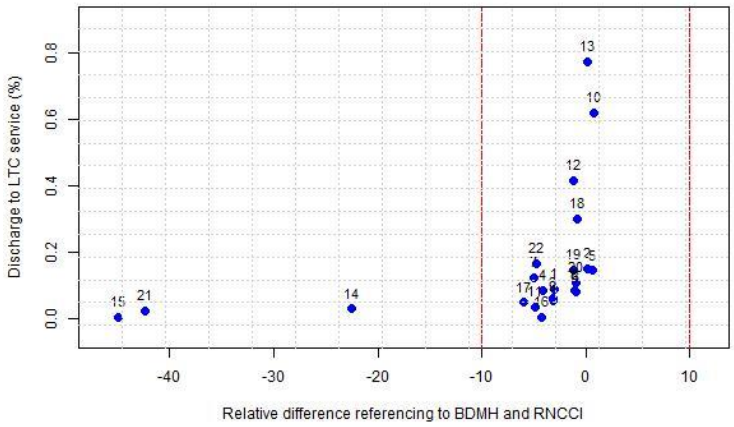


Figure 1. Scatterplot of the public hospitals mainland in Portugal comparing the relative difference of referrals registered in the Gestcare and the LTCD registered in the Hospital Morbidity Database. Hospitals 14,15 and 21 were excluded, hence, the number of hospitals used in this study was 19.

3.1.2. Variables

As described previously, RNCCI includes three levels of nursing care, based on the estimated length of stay and type of care to be provided. In this analysis, a discharge destination to UMLD (DSP=63) was considered as the main outcome.

Other variables considered were age, sex, principal diagnosis, comorbidities (identified using secondary diagnoses), admission type and hospital transfer.

3.2. Data pre-processing

Before the analysis, in order to facilitate data analytic tasks and to have more understandable results, the data was pre-processed. Age was categorised in 5 age groups: <18, 18-49, 50-64, 65-79 and >80.

Due to the huge number of possible different codes for the principal diagnosis (17582 in ICD-9-CM and 93830 in ICD-10-CM), meaning the reason for hospitalisation, they were categorised in 4 different ways considering: (1) 19 and 21 groups from the tabular list of diseases and injuries of ICD-9-CM and ICD-10-CM, respectively, representing the main groups or chapters defined within the ICD hierarchy – ICD-9-CM main cat and ICD-10-CM main cat, respectively; (2) 151 groups for ICD-9-CM and 223 groups for ICD-10-CM representing more detailed categories within the main ICD groups (chapters) ICD-9-CM detailed and ICD-10-CM detailed, respectively; (3) 285 levels from single-level Clinical Classification Software (CCS) – CCS single level; and (4) the CCS level 2 (136 levels) for data coded with ICD-9-CM - CCS lvl2 [22]; 540 levels from CCS Refined (CCSR) for data coded with ICD-10-CM - CCSR [23].

For the comorbidities, the Elixhauser Comorbidity Index [46], [47] which allowed summarizing the secondary diagnosis in the following 30 comorbidities: CHF (Congestive Heart Failure), Arrhythmia, Valvular, PHTN (Pulmonary Circulation Disorders), PVD (Peripheral Vascular Disease), HTN (Hypertension), Paralysis, NeuroOther (Other Neurological Disorders), Pulmonary, DM (Diabetes Mellitus without complications), DMcx (Diabetes Mellitus with complications), Hypothyroid, Renal Failure, Liver, PUD (Peptic Ulcer Disease), Lymphoma, Mets (Metabolic Equivalents), HIV, Tumor, Rheumatic, Coagulopathy, Obesity, Weight Loss, FluidsLytes (Fluid and Electrolyte Disorders), Blood Loss, Anemia, Alcohol, Drugs, Psychoses and Depression.

3.3.Data analysis

Hierarchical clustering techniques were used to analyse and obtain the optimal number of clusters. With hierarchical clustering the data are not divided into a particular number of clusters at a single step. These techniques can be divisive or agglomerative algorithms. Divisive hierarchical algorithms are constructed top-down, starting from a single cluster with all individuals. This cluster is divided into two clusters which are, in turn, split into other subclusters until there are n clusters each containing a single individual. On the other hand, agglomerative hierarchical algorithms are built bottom-up,

starting from n clusters with each object as a different cluster. Then, the closest two clusters merge in new clusters until there is a single cluster containing all individuals [48], [49].

With the objective to determine which hierarchical clustering method fitted better and should be applied to the data, the coefficients of each possible method were calculated and the method with higher coefficient was selected. Thus, the data were analysed with the selected method and to obtain the optimal number of clusters, which is the number of clusters that better distributes data, silhouette and elbow plots were helpful. These methods consist in plotting the value of the clustering criterion against the number of groups. Considering that using silhouette method, the number of clusters that better fits the data is the one with higher Average Silhouette Width. As for the elbow method, it consists in finding an "elbow" in the plot. It is important to note that this approach may be subjective [48].

After determining the number of clusters, Decision Tree methods were applied to classify the data and determine the characteristics of each cluster [50]. Using C5.0 Decision-Tree algorithm, the tree was created according to the data obtained through the hierarchical clustering analysis and the number of clusters was the optimal number of clusters determined.

To all observation in the data a label was given based in the cluster to which it belongs. Then, the data was split in datasets for training (75% of data) and testing (25% of data) to create the model. The training of each trial for every method was evaluated and each trial had an error percentage associated to the number of episodes that were placed in other cluster that was not the correct one. For each of these trials it is known the rules used by the model to place each observation in a cluster and its error percentage. The rules of the trial with the smallest error were obtained to classify each cluster.

Once the principal diagnosis variable was categorised with different approaches, the analysis was applied to all approaches, resulting in different outcomes with different numbers of clusters with different characteristics. These differences were presented and the proportion of inpatient episodes discharged to LTC was calculated for each cluster of the different approaches. These proportions were obtained with the determination of the proportion of episodes with the characteristics of each cluster in the data with LTCD and the number of episodes with the same characteristics in the data with episodes with all destinations after discharge.

Descriptive statistical analyses were performed using Rstudio version 2021.09.1 (RStudio Team, Boston, MA) and R software version 4.1.2.

To evaluate the performance of the developed models, besides obtaining the accuracy, which represents the number of samples classified correctly over a total number of samples, it was also calculated the precision, recall and F1 score of each model. Precision is the probability that an object is relevant given that it is returned by the system and recall can be defined as the probability that a relevant object is returned. As for the F1 score, this metric is the combination of the results for precision and recall [51], [52].

CHAPTER 4 – RESULTS

4. Results

4.1. Descriptive data analysis

Table 1 describes demographic and clinical characteristics of all episodes used in the analysis from the 2012-2016 and 2017 datasets.

Table 1. Distribution of demographic and clinical characteristics: age, sex, mode of admission and hospital transfer.

	2012 to 2016			2017		
	N	**%**	**Mean**	**N**	**%**	**Mean**
Age (years)						
<18	2	0.06		0	0.00	
18-49	209	6.06		52	5.24	
50-64	508	14.72	74.64	161	16.50	74.34
65-79	1176	34.08		371	38.01	
>80	1556	45.09		392	40.16	
Sex						
Male/Female	1728	50.07		485	49.69	
Mode of admission						
Programmed	193	5.59		81	8.30	
Urgent	3258	94.41		895	91.70	
Transfer from other hospital						
Yes	72	2.09		0	0	
No	3379	97.91		976	100	
Total	3451	100		976	100	

Through the descriptive analysis of all hospitalizations with LTCD (Table 2), in both datasets, it is possible to notice an increase in the number of episodes with increasing age and a slight decrease in ages over 90 years, with the average age being 74 years.

On the other hand, in terms of gender distribution, there are no significant differences between the proportion of female and male patients. Therefore, it is not

possible to state that there is a relation between sex and LTC, as seen in (Burton et al., 2018; Chen et al., 2021).

Regarding the admission mode, most hospitalizations were done through emergency services (over 90% of the cases). Furthermore, 2.09% of episodes in the dataset from 2012 to 2016 were transferred from other hospital. For the 2017 dataset, there were no recorded transferences.

In Figure 2 and 3 it can be seen the distribution of the individuals by age and sex for 2012 to 2016 dataset and 2017 dataset, respectively.

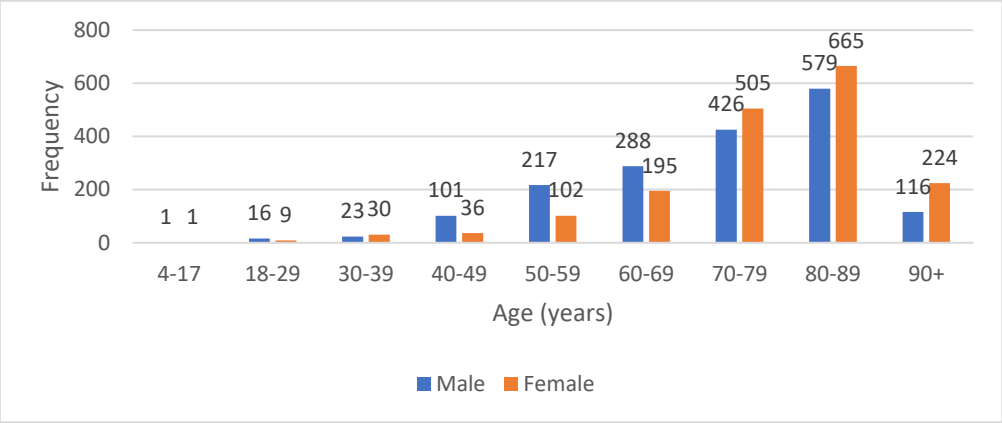


Figure 2. Distribution of individuals of sample from 2012 to 2016 (ICD-9-CM) by age and sex.

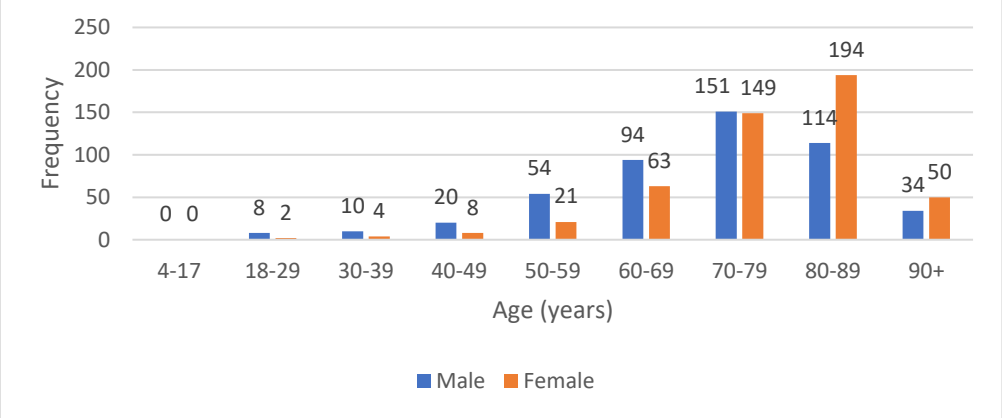


Figure 3. Distribution of individuals of sample from 2017 (ICD-10-CM) by age and sex.

Looking for the distribution by age and sex (Figures 2 and 3), there is an increase in the number of females with the increase of age compared with the number of males with the same age, specially between 80-89 and more than 90 years. Until the ages between 60-69 for the 2012 to 2016 dataset and 70-79 for the 2017 dataset, the number of males in each age range was higher than the number of females.

In order to be determined the most frequent principal diagnoses, which lead to the entry of the patient in the hospital, their frequencies were calculated. The results for both datasets are displayed in Table 2 and Table 3.

Table 2. The most frequent groups of principal diagnoses in 2012 to 2016 dataset.

Principal Diagnoses	Frequency	%
Cerebrovascular disease (430-438)	725	21.0084
Pneumonia and influenza (480-488)	369	10.6925
Other diseases of urinary system (590-599)	331	9.5914
Acute respiratory infections (460-466)	167	4.8391
Supplementary classification of factors influencing health status and contact with health services (v01-v89)	156	4.5204
Fractures (800-829)	135	3.9119
Other diseases of digestive system (570-579)	87	2.5210
Malignant neoplasm of digestive organs and peritoneum (150-159)	86	2.4920
Complications of surgical and medical care, not elsewhere classified (996-999)	80	2.3181
Other forms of heart disease (420-429)	77	2.2312
Pneumoconioses and other lung diseases due to external agents (500-508)	73	2.1153
Other metabolic and immunity disorders (270-279)	60	1.7386
Diseases of arteries, arterioles, and capillaries (440-449)	52	1.5068
Chronic obstructive pulmonary disease and allied conditions (490-496)	52	1.5068
Malignant neoplasm of other and unspecified sites (190-199)	52	1.5068
Diseases of other endocrine glands (249-259)	51	1.4778
Other diseases of respiratory system (510-519)	48	1.3909
Other diagnoses with less frequency	850	24.6305

Table 3. The most frequent groups of principal diagnoses in 2017 dataset.

Principal Diagnoses	Frequency	%
Cerebrovascular diseases I60-I69	208	21.3114
Influenza and pneumonia J09-J18	112	11.4754
Other diseases of the urinary system N30-N39	67	6.8647
Other bacterial diseases A30-A49	51	5.2254
Injuries to the hip and thigh S70-S79	48	4.9180
Factors influencing health status and contact with health services (Z00-Z99)	44	4.5081
Other acute lower respiratory infections J20-J22	39	3.9959
Osteoarthritis M15-M19	21	2.1516
Lung diseases due to external agents J60-J70	19	1.9467
Other forms of heart disease I30-I52	18	1.8442
Injuries to the head S00-S09	18	1.8442
Chronic lower respiratory diseases J40-J47	17	1.7418
Complications of surgical and medical care, not elsewhere classified T80-T88	14	1.4344
Other diagnoses with less frequency	300	30.7377

About the principal diagnoses, in both datasets (Tables 3 and 4), the three most frequent were: “Cerebrovascular diseases” (approximately 21%), “Pneumonia and influenza” (approximately 10%) and “Other diseases of the urinary system” (approximately 9%).

Furthermore, the distribution of the comorbidities summarized by the Elixhauser Comorbidity Index were also obtained and placed in Table 4.

Table 4. Distribution of comorbidities for both datasets.

Dataset	**2012 to 2016**			**2017**		
Comorbidity	No	Yes	% of Yes	No	Yes	% of Yes
Hypertension	1931	1520	44.0452	439	537	55.02049
Arrhythmia	2415	1036	30.02028	699	277	28.38115
Fluid and Electrolyte Disorders	2594	857	24.83338	696	280	28.68852
Diabetes mellitus without complications	2674	777	22.51521	784	192	19.67213
Other Neurological Disorders	2757	694	20.11011	805	171	17.52049
Paralysis	2865	586	16.98059	815	161	16.4959
Congestive Heart Failure	2886	565	16.37207	813	163	16.70082
Anemia	2909	542	15.70559	910	66	6.762295
Renal Failure	3019	432	12.51811	857	119	12.19262
Tumor	3116	335	9.707331	942	34	3.483607
Alcohol	3169	282	8.171544	873	103	10.55328
Pulmonary	3205	246	7.128369	880	96	9.836066
Obesity	3205	246	7.128369	908	68	6.967213
Depression	3211	240	6.954506	908	68	6.967213
Metabolic Equivalents	3250	201	5.824399	945	31	3.17623
Valvular	3241	210	6.085193	932	44	4.508197
Hypothyroid	3289	162	4.694292	928	48	4.918033
Diabetes Mellitus with complications	3304	147	4.259635	895	81	8.29918
Peripheral Vascular Disease	3306	145	4.201681	958	18	1.844262
Liver	3335	116	3.361345	931	45	4.610656
Coagulopathy	3347	104	3.013619	936	40	4.098361
Psychoses	3373	78	2.260214	966	10	1.02459
Pulmonary Circulation Disorders	3407	44	1.274993	959	17	1.741803
Weight Loss	3409	42	1.217039	925	51	5.22541
Rheumatic	3411	40	1.159084	962	14	1.434426
Peptic Ulcer Disease	3410	41	1.188061	976	0	0.000000
Drugs	3418	33	0.956245	968	8	0.819672
Blood Loss	3421	30	0.869313	969	7	0.717213

Lymphoma	3429	22	0.637496	967	9	0.922131
HIV	3451	0	0	972	4	0.409836

Concerning the distribution of comorbidities, the most frequent was hypertension for both datasets (44% and 55%). The most frequent comorbidities also included arrhythmia (approximately 30%), fluid and electrolyte disorders (between 25% and 29%), and diabetes mellitus without complications (between 22% and 18%).

Table 5 describes the demographic and clinical characteristics of all hospital admissions between 2012 and 2017 for all destinations after discharge within the hospitals selected to this study. Among all 1956504 episodes, only 4146 (0.21%) were discharged to LTC.

Table 5. Distribution of the sociodemographic and clinical characteristics of all episodes between 2012 and 2017 with all destinations after discharge within the hospitals selected in this study.

	2012-2017	
	N	**%**
Age (years)		
<18	291429	14.8954
18-49	475883	24.3231
50-64	322299	16.4732
65-79	479037	24.4843
>80	387856	19.8239
Sex		
Male	892204	45.6
Destination after discharge (DSP)		
For home	1734170	88.64
To another institution	62241	3.18
Home service	3971	0.20
Departure against medical opinion	15184	0.78
Specialized aftercare	10392	0.53
Deceased	120384	6.15
Palliative-care	2300	0.12
Post-Hospital Care	3716	0.19
Long-term hospital care	4146	0.21
Mode of admission		
Programmed	515792	26.3629
Urgent	1417077	72.4290
Access	6	0.0003
PECLEC*	7	0.0003
Private Medicine	2	0.0001
SIGIC*	23391	1.1955
PACO*	76	0.0039
External SIGIC	153	0.0078

ICD Version		
ICD-9-CM	1627495	83.18
ICD-10-CM	329009	16.82
Total	1956504	100

*PECLEC - *Programa Especial de Recuperação das Listas de Espera* (Special Program for Recovery of Waiting Lists)

*SIGIC - *Sistema Integrado de Gestão de Inscritos para Cirurgia* (Integrated Surgery Enrollment Management System)

*PACO - *Plano de Acesso à Cirurgia Oftalmológica* (Access Plan for Ophthalmologic Surgery)

The most frequent principal diagnoses in these episodes were "Liveborn Infants According to Type Of Birth" (6.79%), "Pneumonia and Influenza" (4.41%) and "Other Forms of Heat Disease" (3.88%) (Table 6).

Table 6. The most frequent groups of principal diagnoses in all episodes between 2012 to 2017 with all destinations after discharge.

Principal diagnoses	Frequency	%
Liveborn Infants According To Type Of Birth	132881	6.79
Pneumonia And Influenza	86354	4.41
Other Forms Of Heart Disease	75925	3.88
Other Diseases Of Digestive System	74644	3.82
Other Diseases Of Urinary System	68763	3.51
Normal Delivery, And Other Indications For Care In Pregnancy, Labor, And Delivery	67217	3.44
Cerebrovascular Disease	62421	3.19
Ischemic Heart Disease	49820	2.55
Fracture Of Lower Limb	41807	2.14
Complications Mainly Related To Pregnancy	39323	2.01
Other diagnoses with less frequency	1257349	64.18

About the most common comorbidities in these episodes, these are displayed in Table 7. Hypertension (26.69%), Arrhythmia (20.74%) and Pulmonary (17.78%) comorbidities were the three most frequent.

Table 7. Distribution of comorbidities for all episodes between 2012 to 2017 with all destinations after discharge.

Dataset	2012 to 2017		
Comorbidity	No	Yes	% of Yes
Hypertension	15053	5481	26.6923
Arrhythmia	16276	4258	20.7363
Pulmonary	16883	3651	17.7802
Fluid and Electrolyte Disorders	16929	3605	17.5562
Anemia	16937	3597	17.5172
Obesity	16959	3575	17.4101
Depression	17029	3505	17.0692
Other Neurological Disorders	17419	3115	15.1699
Alcohol	17541	2993	14.5758

Renal Failure	17596	2938	14.3079
Congestive Heart Failure	17610	2924	14.2398
Hypothyroid	17794	2740	13.3437
Diabetes Mellitus without complications	17966	2568	12.5060
Liver	18073	2461	11.9850
Coagulopathy	18132	2402	11.6976
Valvular	18152	2382	11.6002
Metabolic equivalents	18558	1976	9.6230
Paralysis	18569	1965	9.5694
Peripheral Vascular Disease	18682	1852	9.0191
Psychoses	18697	1837	8.9461
Diabetes Mellitus with complications	18726	1808	8.8049
Rheumatic	18741	1793	8.7318
Tumor	18777	1757	8.5565
Drugs	18791	1743	8.4883
Peptic Ulcer Disease	18887	1647	8.0208
Lymphoma	19178	1356	6.6036
Pulmonary Circulation Disorders	19301	1233	6.0046
Blood Loss	19405	1129	5.4981
Weight Loss	19536	998	4.8602
HIV	19657	877	4.2709

4.2. Cluster and decision tree analysis

After the descriptive analysis, it was conducted the analysis with clustering and decision tree methods. First, the agglomerative coefficients for the methods *average*, *single*, *complete* and *ward* and the divisive coefficient were calculated for all approaches and both datasets, using the *cluster* package. The values obtained are present in Table 8.

Table 8. Agglomerative and divisive coefficients for each method and dataset.

Dataset and method used to categorise principal diagnosis	**Agglomerative coefficient**				**Divisive coefficient**
	Average method	S*ingle* method	*Complete* method	*Ward* method	
2012-2016 – ICD-9-CM main categories	0.8793	0.6269	0.9334	0.9963	0.9262
2012-2016 – ICD-9-CM detailed categories	0.9751	0.6298	0.9857	0.9993	0.9846
2012-2016 – CCS single-level	0.9879	0.6051	0.9920	0.9997	0.9915

2012-2016 – CCS level 2	0.9780	0.6320	0.9871	0.9994	0.9862
2017 – ICD-10-CM main categories	0.8471	0.5931	0.9234	0.9921	0.9184
2017 – ICD-10-CM detailed categories	0.9694	0.5654	0.9838	0.9980	0.9827
2017 – CCSR	0.9812	0.5955	0.9881	0.9991	0.9874

After analysing the results of each coefficient, it was possible to understand that the hierarchical clustering method with the highest coefficient was the *ward* method from agglomerative clustering, for all datasets. Thus, this method was applied in order to calculate the optimal number of clusters of each dataset using the silhouette and elbow methods. The resulting plots can be found in Figures 4 up to 10.

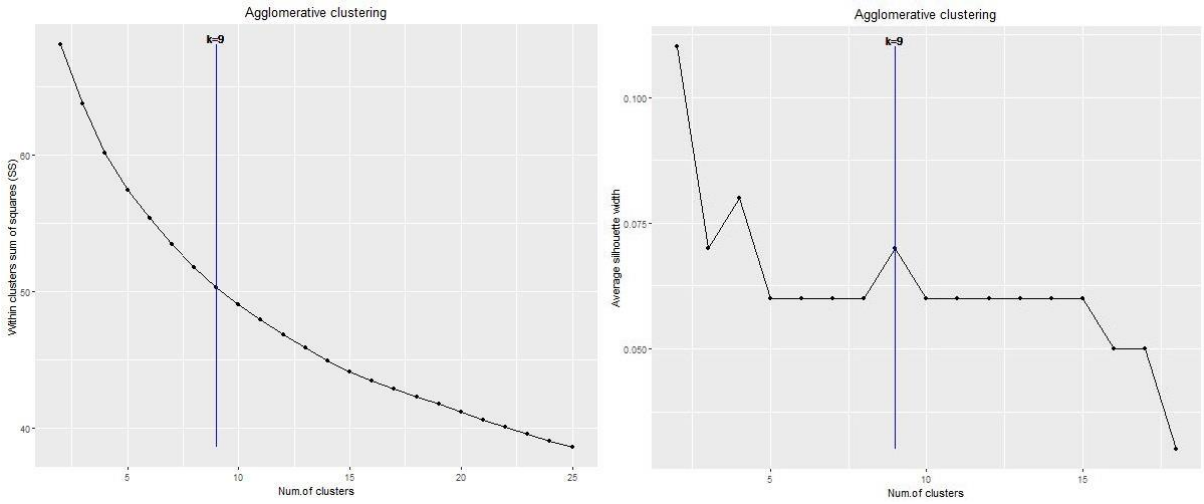


Figure 4. Elbow (on the left) and Silhouette (on the right) plots for 2012-2016 dataset with principal diagnosis categorised with ICD-9-CM main categories and with optimal number of clusters signalled.

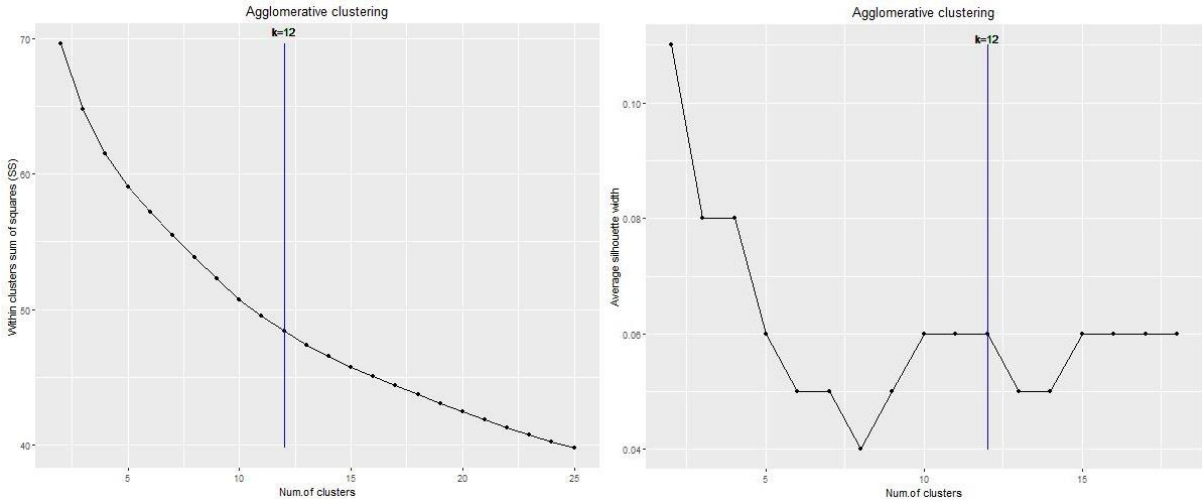


Figure 5. Elbow (on the left) and Silhouette (on the right) plots for 2012-2016 dataset with principal diagnosis categorised with ICD-9-CM detailed categories and with optimal number of clusters signalled.

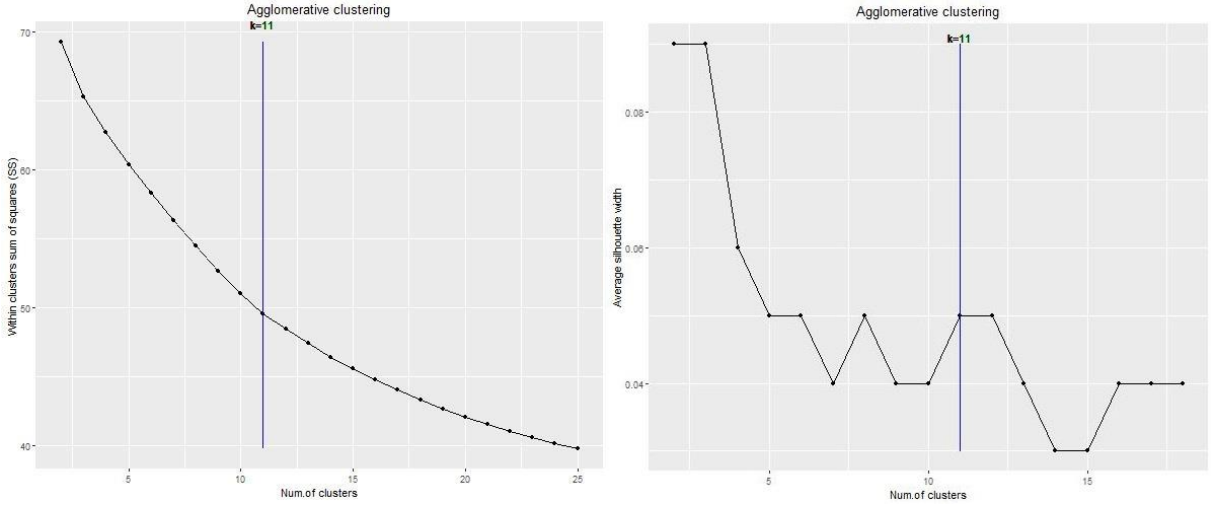


Figure 6. Elbow (on the left) and Silhouette (on the right) plots for 2012-2016 dataset with principal diagnosis categorised with CCS single level and with optimal number of clusters signalled.

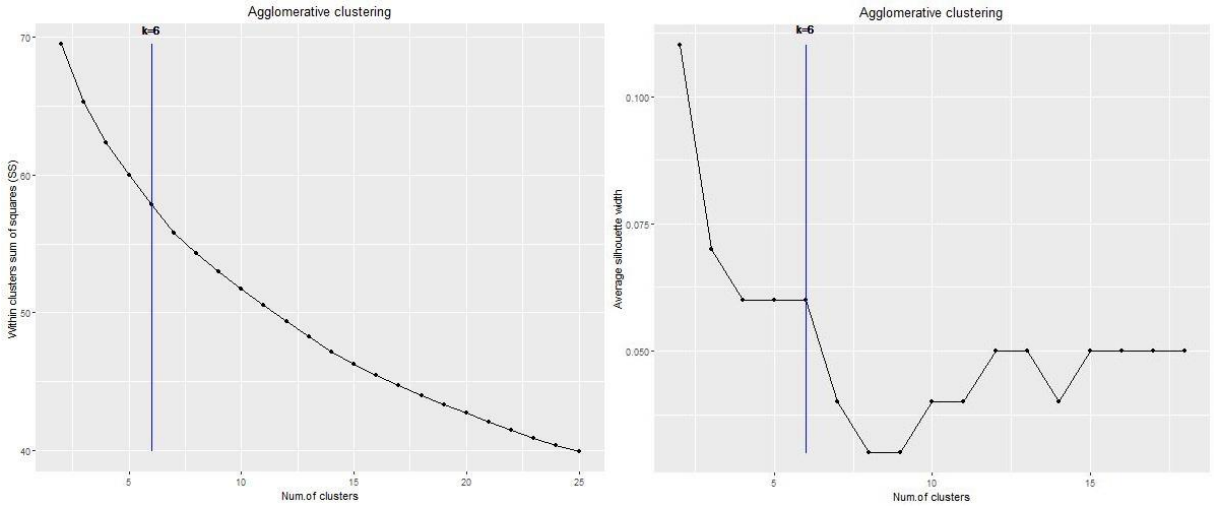


Figure 7. Elbow (on the left) and Silhouette (on the right) plots for 2012-2016 dataset with principal diagnosis categorised with CCS level 2 and with optimal number of clusters signalled.

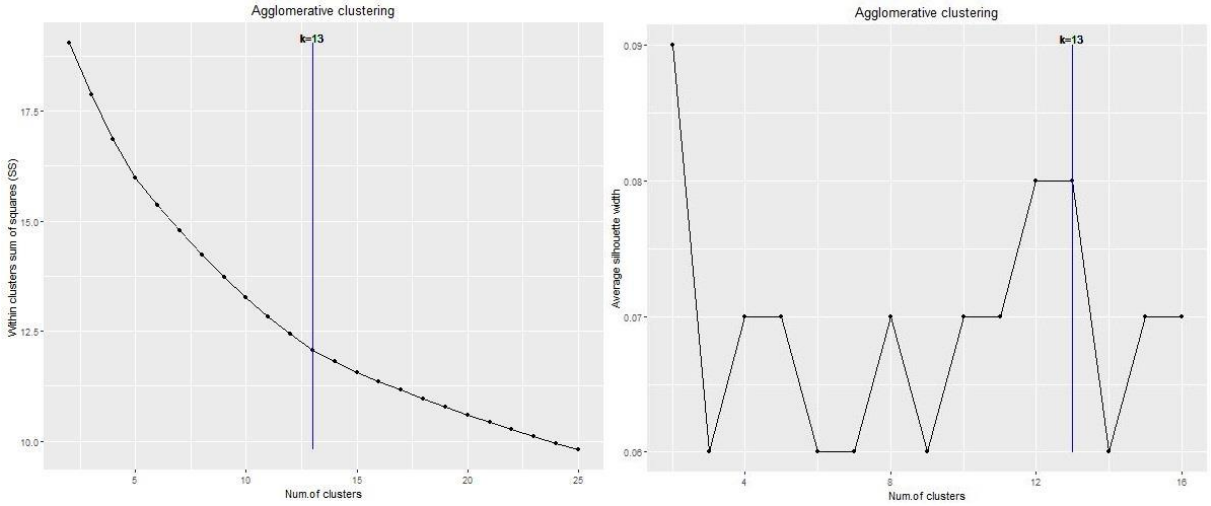


Figure 8. Elbow (on the left) and Silhouette (on the right) plots for 2017 dataset with principal diagnosis categorised with ICD-10-CM main categories and with optimal number of clusters signalled.

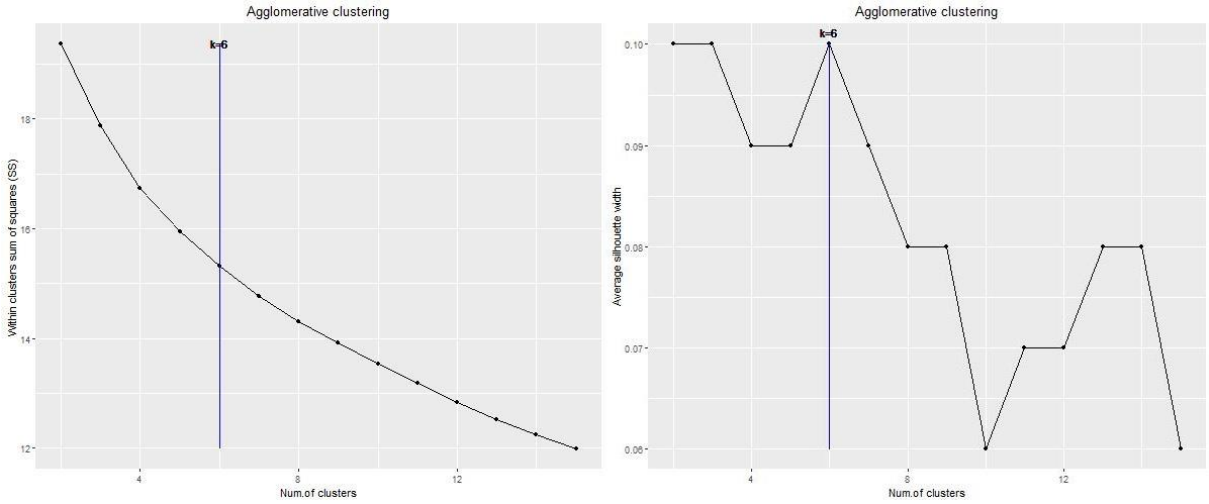


Figure 9. Elbow (on the left) and Silhouette (on the right) plots for 2017 dataset with principal diagnosis categorised with ICD-10-CM detailed categories and with optimal number of clusters signalled

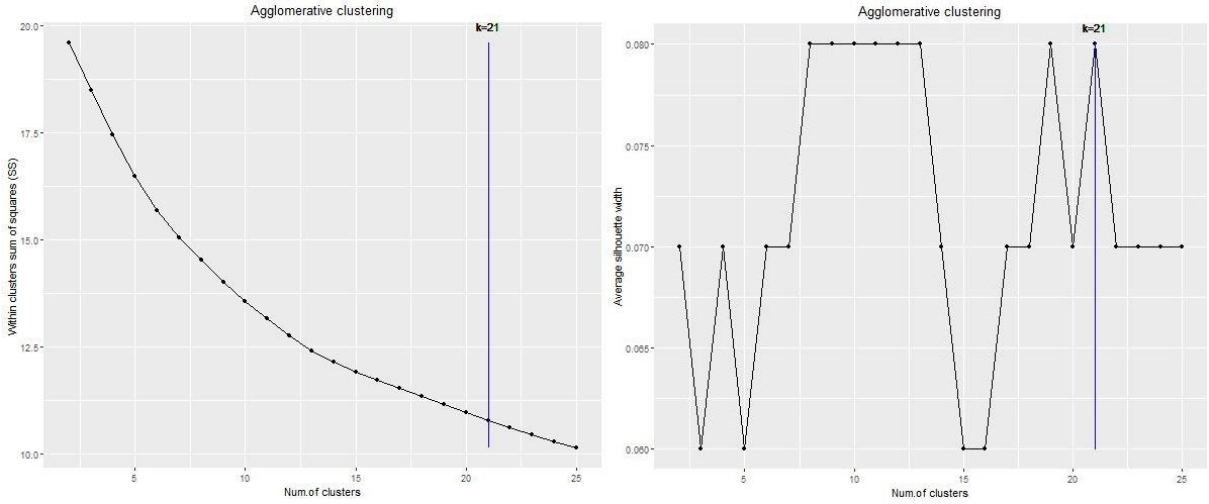


Figure 10. Elbow (on the left) and Silhouette (on the right) plots for 2012-2016 dataset with principal diagnosis categorised with CCSR and with optimal number of clusters signalled.

Once it is important to have a considerable number of clusters in order to obtain better results, analysing the resulting plots, it was possible to understand the optimal number for each dataset, displayed in Table 9.

Table 9. Optimal number of clusters for each dataset.

Dataset and method used to categorise principal diagnosis	Number of principal diagnosis categories	Optimal number of clusters
2012-2016 – ICD-9-CM main categories	19	9
2012-2016 - ICD-9-CM detailed categories	151	12
2012-2016 – CCS single-level	285	11
2012-2016 – CCS level 2	136	6
2017 – ICD-10-CM main categories	21	13
2017 – ICD-10-CM detailed categories	223	6
2017 – CCSR	540	21

Subsequently, the Decision Tree method was applied to each dataset and the models were trained and tested. The model was trained with several trials and for each of these trials it is known its error percentage, according to the number of errors that the model made when placing the episodes in the respective clusters. The errors associated to each of these trials are presented in Figures 11 to 17.

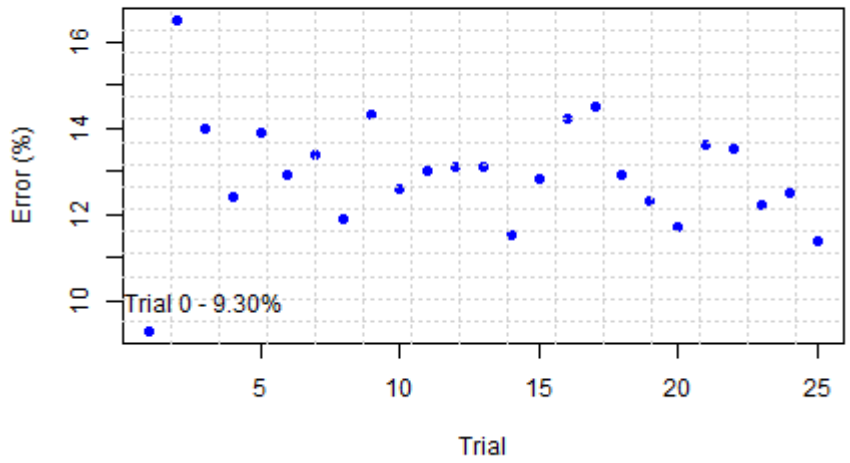


Figure 11. Errors of each trial for the 2012-2016 dataset with principal diagnosis categorised with ICD-9-CM main categories and with the trial with less error signalled.

Regarding the 2012-2016 dataset, analysing Figure 11, for the ICD-9-CM main categories method, the trial with less error associated was Trial 0 with 9.30%.

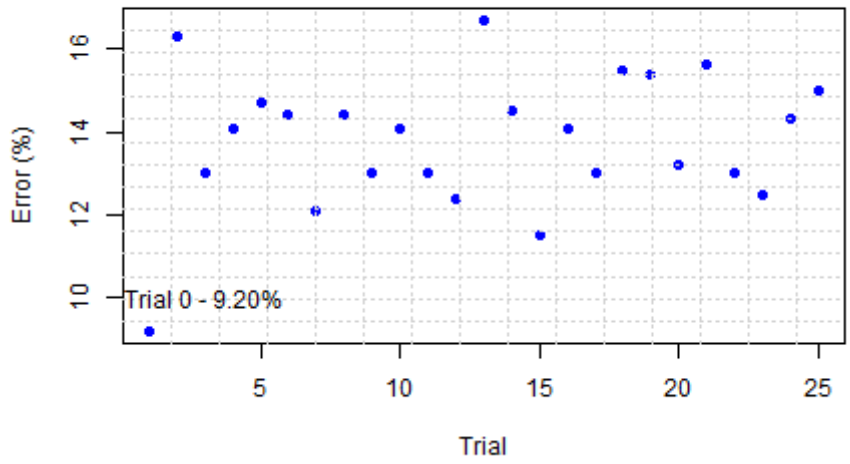


Figure 12. Errors of each trial for the 2012-2016 dataset with principal diagnosis categorised with ICD-9-CM detailed categories and with the trial with less error signalled.

Examining Figure 12, the trial with less error associated for the ICD-9-CM detailed categories method was Trial 0 with 9.20%.

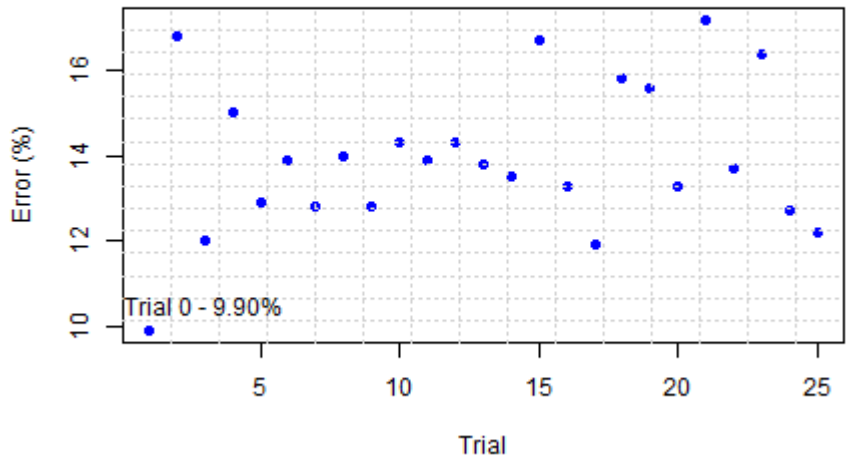


Figure 13. Errors of each trial for the 2012-2016 dataset with principal diagnosis categorised with CCS single level and with the trial with less error signalled.

For the CCS single level method, the trial with less error associated was Trial 0 with 9.90%, as it can be seen in Figure 13.

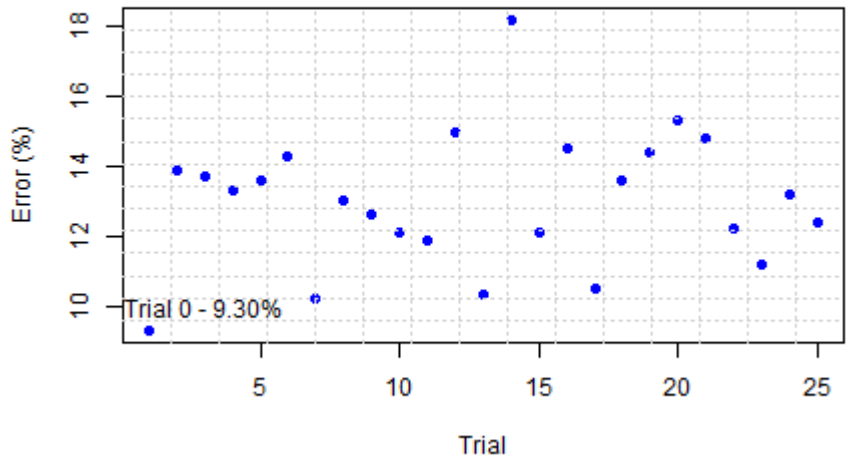


Figure 14. Errors of each trial for the 2012-2016 dataset with principal diagnosis categorised with CCS level 2 and with the trial with less error signalled.

Concerning the CCS level 2 method, the trial with less error associated was Trial 0 with 9.30% (Figure 14).

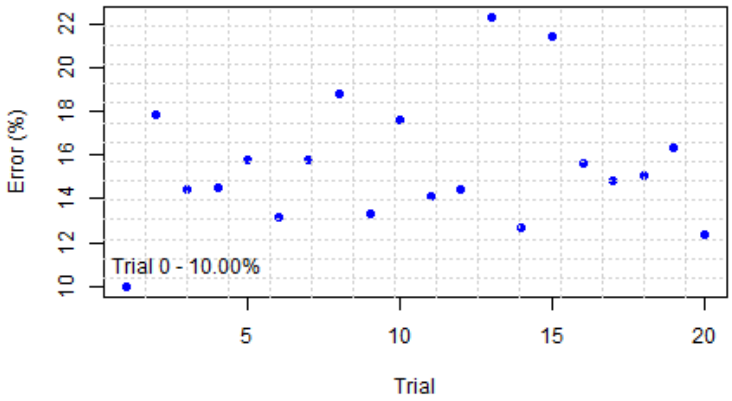


Figure 15. Errors of each trial for the 2017 dataset with principal diagnosis categorised with ICD-10-CM main categories and with the trial with less error signalled.

On the other hand, for the 2017 dataset, analysing Figure 15, the trial with less error associated for the ICD-10-CM main categories method was Trial 0 with 10.00%.

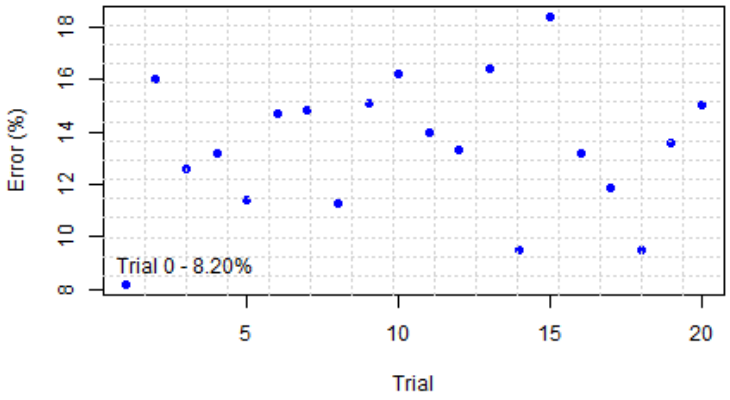


Figure 16. Errors of each trial for the 2017 dataset with principal diagnosis categorised with ICD-10-CM detailed categories and with the trial with less error signalled.

Regarding the ICD-10-CM detailed categories method, the trial with less error associated was Trial 0 with 8.20%, as it is displayed in Figure 16.

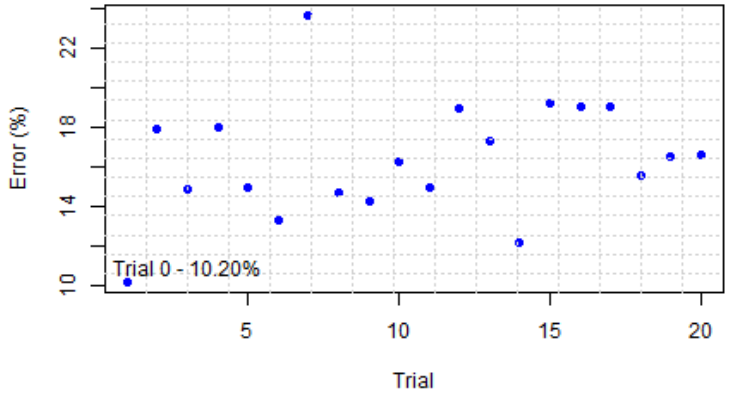


Figure 17. Errors of each trial for the 2017 dataset with principal diagnosis categorised with CCSR and with the trial with less error signalled.

About the CCSR method, the trial with less error associated was Trial 0 with 10.20%, as it is observable in Figure 17.

By analysing the errors of each trial, it was possible to select the trial with less error whose rules would be used to characterize each of the clusters. Tables 10 to 16 display the results and the characteristics associated with each cluster for each dataset.

Table 10. Clusters for the 2012-2016 dataset with principal diagnosis categorised with ICD-9-CM main categories.

Cluster	Age	Sex	Admission type	Transfer	Principal Diagnostic	Comorbidities
1 (328 episodes)	Older than 50	Both	-	-	Diseases of the Genitourinary System; Diseases of the Respiratory System; Neoplasms; Injury and Poisoning; Diseases of the Digestive System; Symptoms signs and ill-defined conditions; Infectious and parasitic diseases	Fluid and Electrolyte Disorders, Pulmonary, Anemia
2 (454 episodes)	Older than 80	Both	Urgent	No	Diseases of the Circulatory System; Diseases of the Digestive System; Diseases of the Genitourinary System; Diseases of the Nervous System and Sense Organs; Injury and Poisoning; Neoplasms; Injury and Poisoning; Infectious and parasitic diseases	Arrhythmia, Paralysis, Hypothyroid
3 (139 episodes)	Between 18 and 79	Both	Urgent	-	Diseases of the Circulatory System; Injury and Poisoning; Mental Disorders	Arrhythmia, Alcohol
4 (415 episodes)	Older than 50	Male	Urgent	-	-	Tumor, Hypothyroid, Peripheral Vascular Disease
5 (718 episodes)	Older than 50	Both	Urgent and Programmed	Yes	Endocrine Nutritional and Metabolic Diseases and Immunity Disorders; Diseases of the Blood and Blood forming organs; Diseases of the Circulatory System; Injury and Poisoning; Supplementary Classification of Factors Influencing Health Status and Contact with Health Services; Diseases of the Digestive System; Diseases of the Genitourinary System	Renal failure, Diabetes Mellitus with complications, Obesity, Peripheral Vascular Disease
6 (141 episodes)	-	-	-	-	-	Other neurological disorders
7 (571 episodes)	Older than 18	Female	Programmed	-	Diseases of the Circulatory System; Diseases of the Digestive System; Neoplasms; Injury and Poisoning; Neoplasms; Diseases of the Genitourinary System; Diseases of the Respiratory System; Endocrine Nutritional and Metabolic Diseases and Immunity Disorders	Obesity, Depression
8 (251 episodes)	Older than 18	Both	-	-	Diseases of the Genitourinary System; Injury and Poisoning; Diseases of the Respiratory System; Diseases of the Nervous System and Sense Organs	Drugs; Other neurological disorders
9 (434 episodes)	Older than 18	Both	Urgent	-	Diseases of the Circulatory System; Diseases of the Genitourinary System; Diseases of the Respiratory System; Diseases of the Blood and Blood forming organs	Depression, Congestive heart failure, Pulmonary circulation disorders, Arrhythmia

Regarding the results of the clustering and decision tree analysis, from the dataset with episodes from 2012 to 2016 with the principal diagnosis categorised with ICD-9-CM main categories (Table 10) resulted 9 different clusters. Cluster 5 is the one with higher number of episodes with 718 male and female patients older than 50 years, with urgent or programmed admissions, transferred from other hospital, with Endocrine Nutritional and Metabolic Diseases and Immunity Disorders, Diseases of the Blood and Blood forming organs, Diseases of the Circulatory System, Injury and Poisoning, Supplementary Classification of Factors Influencing Health Status and Contact with Health Services, Diseases of the Digestive System or Diseases of the Genitourinary System as principal diagnosis and Renal failure, Diabetes Mellitus with complications, Obesity and/or Peripheral Vascular Disease as comorbidities.

Table 11. Clusters for the 2012-2016 dataset with principal diagnosis categorised with ICD-9-CM detailed categories.

Cluster	Age	Sex	Admission type	Transfer	Principal Diagnostic	Comorbidities
1 (355 episodes)	Older than 18	Both	-	-	-	Diabetes Mellitus with complications, Peptic ulcer disease, Diabetes Mellitus without complications
2 (381 episodes)	Between 65 and 79	Female	-	-	-	Other neurological disorders, Arrhythmia, Paralysis, Tumor
3 (211 episodes)	Older than 18	Both	-	-	Poisoning by drugs medicinal and biological substances	Peripheral vascular disease, Diabetes Mellitus with complications, Renal
4 (519 episodes)	Older than 18	Both	Programmed and Urgent	-	Dorsopathies; fractures; Human immunodeficiency virus hiv infection; Infections of skin and subcutaneous tissue; Mental retardation; Open wounds; Other and unspecified effects of external causes; Other.diseases of intestines and peritoneum; other diseases of respiratory system; Other diseases of skin and subcutaneous tissue; Other diseases of urinary system; Other disorders of the central nervous system; Other metabolic and immunity disorders; Pneumoconioses and other lung diseases due to external agents; Supplementary classification of factors influencing health status and contact with health services	Liver, Drugs
5 (243 episodes)	-	Both	Programmed and Urgent	No	Cerebrovascular disease; Acute respiratory infections; Fractures; Ischemic heart disease; Other forms of heart disease; Pneumonia and influenza; Poliomyelitis and other non-arthropod borne viral diseases and prion diseases of central nervous system; Supplementary classification of factors influencing health status and contact with health services	Diabetes Mellitus without complications, Alcohol, Other neurological disorders
6 (169 episodes)	Between 18 and 64 and older than 80	Both	-	-	-	Congestive heart failure, Pulmonary circulation disorders, Valvular, Hypothyroid, Fluid and Electrolyte Disorders
7 (171 episodes)	Older than 65	Female	-	-	-	Fluid and Electrolyte Disorders, Obesity
8 (299 episodes)	Older than 18	Both	Urgent	-	-	Tumor, Pulmonary
9 (536 episodes)	Older than 18	Female	-	Yes	-	Obesity, Depression
10 (177 episodes)	Between 18 and 49 and older than 65	-	-	-	-	Anemia
11 (264 episodes)	Older than 80	Both	-	-	-	Arrhythmia, Hypertension, Congestive heart failure, Anemia, Hypothyroid
12 (126 episodes)	-	-	-	-	-	Metabolic equivalents

As for the results of the dataset with episodes from 2012 to 2016 and with the principal diagnosis categorised with ICD-9-CM detailed categories (Table 11), these

showed 12 different clusters. The cluster with higher number of episodes is Cluster 9 with 536 females older than 18 years, transferred from other hospital and with Obesity and/or Depression as comorbidities.

Table 12. Clusters and characteristics for the 2012-2016 dataset with principal diagnosis categorised with CCS single level.

Cluster	Age	Sex	Admission type	Transfer	Principal Diagnostic	Comorbidities
1 (501 episodes)	Between 18 and 79	Both	Programmed	-	-	Liver
2 (315 episodes)	-	Male	Urgent	-	Acute cerebrovascular disease; Aspiration pneumonitis; Back problem; Brain injury; Chronic obstructive pulmonary disease; HIV infection, Intestinal obstruction; Intracranial injury; Other aftercare; Other benign neoplasm;, Other hereditary Central Nervous System disease; Other hereditary Central Nervous System infection; Other Central Nervous System disease; Other cerebrovascular disease; Other Gastrointestinal disease; Pathological fracture; Spinal cord injury; Spontaneous abortion; Ulcer skin; Urinary Tract Infection	Paralysis, Other neurological disorders, Alcohol
3 (252 episodes)	Older than 50	Female	Programmed and Urgent	No	Acute cerebrovascular disease; Congestive heart failure; Encephalitis; Leg fracture; Other nutrition disease; Other liver disease; Septicemia; Urinary Tract Infection	Hypertension, Other neurological disorders
4 (312 episodes)	Older than 18	Both	-	-	-	Renal, Arrhythmia, Hypothyroid, Peripheral vascular disease, Fluid and Electrolyte Disorders, Diabetes Mellitus with complications, Valvular
5 (229 episodes)	Between 50 and 79	Both	-	No	Bronchitis; Coronary atherosclerosis; Pneumonia; Septicemia	Other neurological disorders, Tumor
6 (214 episodes)	Older than 18	Female	-	-	Delirium dementia amnestic other cognitive; Fracture of neck of femur (hip)	Pulmonary circulation disorders, Valvular, Congestive heart failure
7 (300 episodes)	Older than 65	Female	Urgent	-	-	Arrhythmia, Hypertension, Paralysis, Pulmonary circulation disorders, Obseity, Tumor
8 (425 episodes)	Older than 18	Male	-	No	-	Tumor, Depression, Arrhythmia, Pulmonary
9 (418 episodes)	-	Female	Urgent	-	-	Depression, Hypertension, Metabolic equivalents
10 (284 episodes)	Older than 18	Both	-	Yes	-	Diabetes Mellitus without complications, Fluid and Electrolyte Disorders
11 (201 episodes)	Older than 50	Both	-	-	Gangrene; Other aftercare; Peripheral atherosclerosis; Ulcer skin; Urinary Tract Infection	Congestive heart failure, Anemia, Hypertension

On the other hand, the results of the same data with principal diagnosis categorised with CCS single level (Table 12) presented 11 clusters. In this analysis, the cluster with higher number of episodes is Cluster 1 with 501 males and females with ages between 18 and 79 years, programmed admission and liver comorbidities.

Table 13. Clusters and characteristics for the 2012-2016 dataset with principal diagnosis categorised with CCS level 2.

Cluster	Age	Sex	Admission type	Transfer	Principal Diagnostic	Comorbidities
1 (962 episodes)	Between 18 and 79	Both	Programmed	-	Delirium, dementia and amnestic and other cognitive disorders; Diabetes mellitus with complications; Other inflammatory condition of skin; Viral infection	Metabolic equivalents, Alcohol, Liver, Pulmonary
2 (449 episodes)	Between 65 and 79	-	-	No	Cerebrovascular disease	Congestive heart failure, Other neurological disorders, Paralysis, Arrhythmia
3 (868 episodes)	-	Both	Urgent	Yes	Anemia; Chronic obstructive pulmonary disease and bronchiectasis; Diseases of the heart; Respiratory infections	Diabetes Mellitus without complications, Diabetes Mellitus with complications, Depression, Peripheral Vascular Disease, Pulmonary Circulation Disorders, Coagulopathy
4 (507 episodes)	Older than 18	Both	-	No	Aspiration pneumonitis food vomitus; Cancer of bronchus lung; Cancer of lymphatic and hematopoietic tissue; Cancer other primary; Central nervous system infection; Diabetes mellitus with complications; Diseases of the heart; Hypertension; Intracranial injury; Other lower respiratory disease; Viral infection; Diseases of the urinary system	Fluid and Electrolyte Disorders, Hypothyroid, Pulmonary Circulation Disorders
5 (263 episodes)	Older than 18	Male	-	No	Aspiration pneumonitis food vomitus; Cerebrovascular disease; Diseases of the urinary system; Epilepsy convulsions; Factors influencing health care; Lower gastrointestinal disorders; Respiratory infections; Bacterial infection; Biliary tract disease; Delirium dementia and amnestic and other cognitive disorders; Diseases of male genital organs; Fluid and electrolyte disorders; Other congenital anomalies; Poisoning; Respiratory failure insufficiency arrest adult; Symptoms signs and ill defined conditions	Other neurological disorders
6 (402 episodes)	Older than 18	Both	-	No	-	Diabetes Mellitus without complications, Peripheral Vascular Disease

As for the results of the same data with principal diagnosis categorised with CCS level 2 (Table 13), these demonstrated 6 different clusters. Cluster 1 was the cluster with higher number of episodes with 962 males and females with ages between 18 and 79 years, programmed admission, with Delirium, Dementia and amnestic and other cognitive disorders, Diabetes mellitus with complications, Other inflammatory condition of skin or Viral infection as principal diagnosis and Metabolic equivalents, Alcohol, Liver and/or Pulmonary as comorbidities.

Table 14. Clusters and characteristics for the 2017 dataset with principal diagnosis categorised with ICD-10-CM main categories.

Cluster	Age	Sex	Admission type	Principal Diagnostic	Comorbidities
1 (63 episodes)	Older than 80	Male	Programmed and Urgent	Diseases of the Circulatory System; Diseases of the Digestive System; Symptoms, Signs and Abnormal Clinical and Laboratory Findings not elsewhere Classified	Arrhythmia, Diabetes Mellitus, Congestive Heart Failure
2 (138 episodes)	Older than 65	Female	Urgent	-	Hypertension, Hypothyroid, Diabetes Mellitus with

					complications, Renal failure
3 (79 episodes)	Older than 50	Both	Programmed and Urgent	Diseases of the Circulatory System; Diseases of the Nervous System	Arrhythmia, Paralysis, Congestive Heart Failure, Other neurological disorders
4 (90 episodes)	Between 65 and 79	Female	Programmed and Urgent	Diseases of the Genitourinary System	Arrhythmia, Obesity, Valvular, Hypertension
5 (87 episodes)	Between 65 and 79	Male	Urgent	-	Diabetes Mellitus
6 (55 episodes)	Between 18 and 64	Male	Urgent	Certain Infectious and Parasitic Diseases; Endocrine Nutritional and Metabolic Diseases	Liver, Alcohol, Fluid and Electrolyte Disorders, Coagulopathy
7 (121 episodes)	Older than 65	Male	Urgent	-	Renal failure, Fluid and Electrolyte Disorders
8 (38 episodes)	-	Both	Urgent	-	Other neurological disorders, Fluid and Electrolyte Disorders
9 (55 episodes)	Older than 80	Both	Urgent	Diseases of the Respiratory System	Pulmonary, Renal failure, Congestive Heart Failure, Tumor
10 (80 episodes)	Between 65 and 79	Female	-	-	Congestive Heart Failure, Fluid and Electrolyte Disorders, Weight Loss
11 (51 episodes)	-	-	Programmed	-	-
12 (71 episodes)	Between 18 and 79	Male	Programmed and Urgent	-	Hypertension, Diabetes Mellitus with complications, Paralysis
13 (48 episodes)	Older than 80	Both	Urgent	-	Arrhythmia, Diabetes Mellitus, Fluid and Electrolyte Disorders, Hypertension, Renal failure, Other neurological disorders

About the results of the dataset from 2017 with principal diagnosis categorised with ICD-10-CM main categories (Table 14), these revealed 13 clusters. The cluster with higher number of episodes was Cluster 2 with 138 episodes, females older than 65 years, urgent admission and with hypertension, hypothyroid, Diabetes Mellitus with complications and/or renal failure comorbidities.

Table 15. Clusters and characteristics for the 2017 dataset with principal diagnosis categorised with ICD-10-CM detailed categories.

Cluster	Age	Sex	Admission type	Principal Diagnostic	Comorbidities
1 (193 episodes)	Older than 50	Male	Urgent	-	Fluid and Electrolyte Disorders, Diabetes Mellitus without complications, Coagulopathy
2 (336 episodes)	Between 50 and 64 and older than 80	Both	Programmed	Malignant neoplasms of ill-defined other secondary and unspecified sites; Influenza and pneumonia; Injuries to the thorax; Injuries to the hip and thigh	Congestive heart failure
3 (130 episodes)	-	Female	-	-	Arrhythmia, peripheral vascular disease, Paralysis
4 (112 episodes)	-	Female	Programmed	-	Pulmonary circulation disorders, Valvular

5 (156 episodes)	-	Male	-	-	Anemia, Liver, Alcohol, Drugs
6 (49 episodes)	Between 65 and 79	Both	-	-	Diabetes Mellitus with complications, Renal, Liver, BloodLoss

The results of the dataset with episodes from 2017 and with the principal diagnosis categorised with ICD-10-CM detailed categories (Table 15) showed 6 different clusters. Cluster 2 was the one with higher number of episodes with 336 males and females between 50 and 64 and older than 80 years, programmed admission, malignant neoplasms of ill-defined other secondary and unspecified sites, influenza and pneumonia, injuries to the thorax and injuries to the hip and thigh as principal diagnosis and congestive heart failure comorbidities.

Table 16. Clusters and characteristics for the 2017 dataset with principal diagnosis categorised with CCSR.

Cluster	Age	Sex	Admission type	Principal Diagnostic	Comorbidities
1 (45 episodes)	-	Male	-	-	-
2 (82 episodes)	Older than 18	Female	Urgent	Nutritional anemia; Other aftercare encounter; Other specified and unspecified lower respiratory disease; Pneumonia except that caused by tuberculosis	-
3 (93 episodes)	Older than 80	Female	Urgent	-	Hypertension
4 (72 episodes)	-	Female	Urgent	Acute hemorrhagic cerebrovascular disease; Cerebral infarction	Hypertension, Other neurological disorders, Paralysis
5 (34 episodes)	Between 50 and 64 and older than 80	Both	-	-	Arrhythmia, Hypertension, Fluid and Electrolyte Disorders, Other neurological disorders, peripheral vascular disease, Coagulopathy
6 (73 episodes)	Between 50 and 79	Male	Urgent	-	Diabetes Mellitus without complications, Hypertension, Fluid and Electrolyte Disorders
7 (40 episodes)	Between 50 and 64	Male	-	-	Hypertension, Alcohol
8 (89 episodes)	Between 65 and 79	Male	Urgent	Fluid and electrolyte disorders; Gastrointestinal and biliary perforation; Heart failure; Nervous system cancers brain; Other aftercare encounter; Peripheral and visceral vascular disease; Pneumonia except that caused by tuberculosis; Pressure ulcer of skin, Sequela of cerebral infarction and other cerebrovascular disease; Spinal cord injury SCI subsequent encounter	Alcohol, Paralysis,
9 (52 episodes)	Between 18 and 79	Female	Urgent	-	Other neurological disorders, Fluid and Electrolyte Disorders, WeightLoss
10 (37 episodes)	Older than 50	Both	-	Cerebral infarction; Pneumonia except that caused by tuberculosis	Renal, Fluid and Electrolyte Disorders, Metabolic equivalents, WeightLoss
11	Between 65 and 79	Female	Urgent	-	Arrhythmia, Pulmonary circulation disorders,

(27 episodes)					Hypertension, Renal, Obesity, Congestive heart failure
12 (45 episodes)	-	Both	Programmed	-	Other neurological disorders, Diabetes Mellitus without complications
13 (23 episodes)	Older than 50	Both	-	-	Anemia
14 (55 episodes)	Older than 80	Female	Programmed and Urgent	-	Fluid and Electrolyte Disorders, Valvular
15 (58 episodes)	-	Female	Urgent	-	Arrhythmia
16 (23 episodes)	Between 65 and 79	Both	Urgent	-	Arrhythmia, Diabetes Mellitus with complications, Congestive heart failure, Renal, Fluid and Electrolyte Disorders, Other neurological disorders
17 (28 episodes)	Between 65 and 79	Both	Urgent	Aspiration pneumonitis; Cerebral infarction; Urinary tract infections	Arrhythmia, Paralysis, Fluid and Electrolyte Disorders
18 (43 episodes)	Older than 80	Male	-	-	Renal, Congestive heart failure
19 (33 episodes)	Between 18 and 79	Male	-	-	Hypertension
20 (20 episodes)	Between 18 and 49	Male	-	-	Alcohol, Drugs, Coagulopathy
21 (4 episodes)	Between 18 and 49	-	-	-	Arrhythmia, Pulmonary, Fluid and Electrolyte Disorders, Hypertension

The results with the same data but with the principal diagnosis categorised with CCSR (Table 16) showed 21 clusters. The cluster with higher number of episodes is Cluster 3 with 93 episodes, females older than 80 years, urgent admission and with hypertension as comorbidity.

Figure 18 displays the distribution of clusters for the proportion of inpatients episodes discharged to LTC, by dataset and by principal diagnosis categorisation method. The clusters signalized are the ones with higher proportion of LTCD in each dataset.

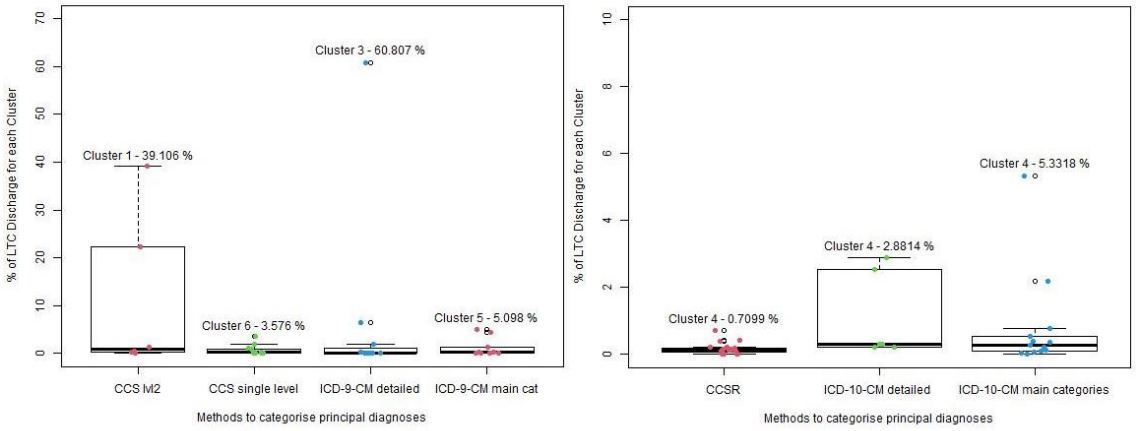


Figure 18. Distribution of clusters for the proportion of inpatient episodes with discharge to LTC, by dataset and principal diagnoses categorisation, in Portugal between 2012 and 2017

Analysing Figure 18, it is possible to see that for the 2012 to 2016 dataset with principal diagnosis categorised with CCS level 2, the cluster with higher LTCD was Cluster 1(39.1%): males and females with ages between 18 and 79, programmed admission, delirium, dementia and amnestic and other cognitive disorders, Diabetes mellitus with complications, other inflammatory condition of skin or viral infection as principal diagnostic and metabolic equivalents, alcohol, liver or pulmonary comorbidities. About the dataset with principal diagnosis categorised with CCS single level, the cluster with higher LTCD was Cluster 6 (3.6%): females older than 18 years with delirium dementia amnestic other cognitive or fracture of neck of femur (hip) as principal diagnosis and pulmonary circulation disorders, valvular or congestive heart failure comorbidities. For the same dataset with principal diagnosis categorised with ICD-9-CM categories detailed, the cluster with higher LTCD was Cluster 3 (60.8%): males and females older than 18 years with poisoning by drugs medicinal and biological substances as principal diagnostic and with peripheral vascular disease, Diabetes Mellitus with complications or renal failure as comorbidities. For the one categorised with ICD-9-CM main categories, the cluster with higher LTCD was Cluster 5 (5.1%): males and females older than 50 years, urgent and programmed admission, transferred from other hospital, with Endocrine Nutritional and Metabolic Diseases and Immunity Disorders, Diseases of the Blood and Blood forming organs, Diseases of the Circulatory System, Injury and Poisoning, Supplementary Classification of Factors Influencing Health Status and Contact with Health Services, Diseases of the Digestive System or Diseases of the

Genitourinary System as principal diagnosis and with renal failure, Diabetes Mellitus with complications, obesity or Peripheral Vascular Disease comorbidities.

As for the 2017 dataset with principal diagnosis categorised with CCSR, the cluster with higher proportion of LTCD was Cluster 4 (0.7%): females with urgent admission, acute haemorrhagic cerebrovascular disease or cerebral infarction as principal diagnosis and hypertension, other neurological disorders or paralysis comorbidities. For the same dataset with principal diagnosis categorised with ICD-10-CM detailed categories, the cluster with higher proportion of LTCD within all episodes was Cluster 4 (2.9%) whose characteristics were: females with programmed admission and pulmonary circulation disorders or valvular comorbidities. About the dataset with principal diagnosis categorised with ICD-10-CM main categories, the cluster with higher number of episodes was Cluster 4 (5.3%): females between 65 and 79 years, programmed and urgent admissions, diseases of the genitourinary system as principal diagnostic and arrhythmia, obesity, valvular or hypertension comorbidities.

It is possible to verify that the cluster with the highest proportion of inpatient episodes discharged to LTC ranged between 3.6% (CCS single level) and 60.8% (ICD-9-CM detailed) for the first dataset, while it ranged between 0.7% (CCSR) and 5.3% (ICD-10-CM main categories) in the second dataset. There is a high level of heterogeneity using different principal diagnosis categorisation, with the highest proportion being obtained using the “ICD-9-CM detailed” option for the first dataset. In this method, the Cluster 3 showed a proportion of 60.8% (211 out of 347) of inpatient episodes being discharged to LTC. This cluster included inpatients older than 18 years with poisoning by drugs medicinal and biological substances as the principal diagnosis and having peripheral vascular disease, diabetes mellitus with complications and/or renal failure as comorbidities. However, for the second dataset, the highest proportion was obtained using the “ICD-10-CM main categories”, which included females between 65 and 79 years-old having a principal diagnosis of "diseases of the genitourinary system" and arrythmia, obesity, valvular and/or hypertension as comorbidities.

Given the high level heterogeneity of the results and that there were some variables that were more used to classify clusters than others, the attribute usage of each method was analysed with the objective to understand these different results between the different approaches. The results of these analysis can be seen in Tables 17 and 18.

Table 17. Attribute usage for each method applied to ICD-9-CM data.

ICD-9-CM main categories		ICD-9-CM detailed categories		CCS single level		CCS level 2		Mean of all methods	
%	Attribute	%	Attribute	Attribute	Attribute	%	Attribute	%	Attribute
100.00%	HTN	100.00%	Sex	100.00%	Arrhythmia	100.00%	Paralysis	99.99%	Paralysis
100.00%	Liver	100.00%	Arrhythmia	100.00%	Paralysis	100.00%	NeuroOther	99.99%	Renal
100.00%	Mets	100.00%	HTN	100.00%	Renal	100.00%	FluidsLytes	99.82%	FluidsLytes
100.00%	FluidsLytes	100.00%	Paralysis	100.00%	Anemia	100.00%	Anemia	99.71%	NeuroOther
99.97%	Renal	100.00%	Renal	99.97%	NeuroOther	100.00%	Alcohol	99.54%	Sex
99.94%	Paralysis	100.00%	Mets	99.91%	CHF	99.97%	Sex	99.50%	Age
99.91%	CHF	99.77%	Age	99.83%	FluidsLytes	99.97%	Renal	99.10%	Arrhythmia
99.83%	Age	99.74%	NeuroOther	99.51%	Age	98.90%	Age	98.41%	Anemia
99.77%	Sex	99.59%	Anemia	98.41%	Sex	96.73%	Arrhythmia	97.99%	CHF
99.65%	Arrhythmia	99.45%	FluidsLytes	98.09%	Mets	95.39%	CHF	97.39%	HTN
99.65%	DM	96.75%	CHF	98.06%	DM	94.67%	DM	97.13%	Mets
99.13%	NeuroOther	95.71%	DM	96.84%	HTN	92.73%	HTN	97.02%	DM
98.73%	Principal diagnose	94.90%	Tumor	94.44%	Tumor	92.67%	Valvular	93.00%	Tumor
97.31%	Alcohol	90.06%	Principal diagnose	87.97%	Pulmonary	92.55%	Tumor	86.63%	Pulmonary
94.06%	Anemia	89.08%	Pulmonary	87.54%	Obesity	91.31%	Hypothyroid	86.58%	Principal diagnose
93.16%	Obesity	84.99%	Obesity	82.38%	Depression	90.41%	Mets	85.67%	Alcohol
92.15%	Hypothyroid	84.38%	Valvular	81.86%	Admission type	89.60%	Pulmonary	83.09%	Depression
90.12%	Tumor	80.38%	Admission type	80.06%	Principal diagnose	87.66%	Coagulopathy	82.38%	Valvular
87.11%	Depression	76.24%	Depression	78.44%	Valvular	86.64%	Depression	82.11%	Hypothyroid
79.86%	Pulmonary	74.21%	PVD	76.73%	Hypothyroid	84.76%	DMcx	80.05%	Obesity
74.96%	DMcx	73.72%	Alcohol	71.83%	PVD	77.48%	Principal diagnose	77.58%	Liver
74.04%	Valvular	69.75%	Liver	71.63%	Alcohol	76.91%	PVD	75.72%	Admission type
73.11%	PVD	68.24%	Hypothyroid	68.59%	Liver	71.98%	Liver	74.02%	PVD
71.60%	Admission type	67.66%	DMcx	67.52%	DMcx	69.05%	Admission type	73.73%	DMcx
64.65%	Coagulopathy	62.10%	PHTN	44.45%	Transfer	55.69%	Psychoses	59.61%	Coagulopathy
63.34%	Psychoses	54.13%	Coagulopathy	44.16%	PHTN	54.51%	Obesity	52.00%	PHTN
54.16%	Transfer	51.32%	WeightLoss	42.60%	PUD	52.91%	PHTN	51.02%	Psychoses
48.83%	PHTN	43.12%	Psychoses	41.93%	Psychoses	46.60%	PUD	46.26%	Transfer
38.71%	Drugs	40.51%	Transfer	31.99%	Coagulopathy	45.90%	Transfer	33.62%	PUD
28.51%	WeightLoss	24.28%	PUD	16.78%	Drugs	25.38%	BloodLoss	26.76%	WeightLoss
21.01%	PUD	22.63%	Drugs	16.46%	Lymphoma	17.15%	Rheumatic	23.67%	Drugs
16.46%	BloodLoss	19.07%	Rheumatic	13.74%	Rheumatic	16.57%	Drugs	14.72%	Rheumatic
8.92%	Rheumatic	10.00%	BloodLoss	11.13%	WeightLoss	16.08%	WeightLoss	14.53%	BloodLoss
7.59%	Lymphoma	5.27%	Lymphoma	6.26%	BloodLoss	10.06%	Lymphoma	9.85%	Lymphoma

Table 18. Attribute usage for each method applied to ICD-10-CM data.

ICD-10-CM main categories		ICD-10-CM detailed categories		CCSR		Mean of all methods	
%	Attribute	%	Attribute	%	Attribute	%	Attribute
100.00%	Sex	100.00%	Sex	100.00%	Sex	100.00%	Sex
100.00%	Admission type	100.00%	Paralysis	100.00%	Admission type	99.42%	Arrhythmia
100.00%	CHF	100.00%	DMcx	100.00%	Arrhythmia	99.42%	Paralysis
100.00%	Arrhythmia	99.80%	Age	100.00%	Liver	99.28%	Liver
100.00%	Paralysis	98.26%	Arrhythmia	100.00%	Alcohol	96.35%	CHF
100.00%	DM	97.85%	Liver	99.69%	HTN	96.04%	Age
100.00%	Liver	93.75%	CHF	99.08%	Paralysis	95.05%	FluidsLytes
100.00%	FluidsLytes	92.93%	HTN	98.87%	FluidsLytes	94.26%	HTN
100.00%	Alcohol	92.52%	Pulmonary	97.54%	DM	94.16%	Alcohol
97.13%	Age	86.27%	FluidsLytes	95.29%	CHF	89.17%	Admission type
94.67%	NeuroOther	82.48%	Alcohol	92.62%	NeuroOther	88.49%	NeuroOther
90.16%	HTN	81.56%	Renal	91.19%	Age	85.59%	DM
88.52%	Principal diagnose	79.51%	Depression	89.55%	Anemia	83.95%	Pulmonary
85.66%	Renal	78.18%	NeuroOther	79.71%	Pulmonary	76.68%	Renal
79.61%	Pulmonary	78.07%	Obesity	75.82%	DMcx	76.30%	DMcx
77.25%	WeightLoss	74.59%	Valvular	64.04%	Valvular	69.30%	Anemia
67.11%	Coagulopathy	67.52%	Admission type	62.91%	Principal diagnose	68.44%	Principal diagnose
63.83%	Depression	63.42%	Anemia	62.81%	Renal	63.80%	Valvular
60.66%	Obesity	59.32%	Mets	46.21%	Coagulopathy	60.42%	Depression
54.92%	Anemia	59.22%	DM	45.29%	PHTN	59.08%	Obesity
53.07%	DMcx	53.89%	Principal diagnose	38.52%	Obesity	53.24%	Coagulopathy
52.77%	Valvular	46.41%	Coagulopathy	38.52%	WeightLoss	49.93%	WeightLoss
51.33%	Tumor	39.55%	PVD	37.91%	Depression	40.27%	PHTN
39.65%	PHTN	35.86%	PHTN	35.14%	Hypothyroid	34.90%	Tumor
30.53%	Hypothyroid	34.02%	WeightLoss	30.64%	Drugs	34.46%	Mets
29.00%	Mets	25.72%	Tumor	27.66%	Tumor	28.28%	Hypothyroid
11.27%	PVD	25.31%	Rheumatic	21.93%	PVD	27.67%	Drugs
6.76%	Lymphoma	24.69%	Drugs	15.06%	Mets	24.25%	PVD
6.15%	Rheumatic	23.98%	BloodLoss	6.76%	HIV	12.09%	BloodLoss
4.51%	HIV	19.16%	Hypothyroid	6.45%	Lymphoma	10.52%	Rheumatic
2.46%	Psychoses	12.19%	Psychoses	2.15%	Psychoses	6.18%	Lymphoma
0.00%	BloodLoss	5.33%	Lymphoma	0.20%	BloodLoss	5.64%	HIV
0.00%	Drugs	0.00%	HIV	0.10%	Rheumatic	5.60%	Psychoses

In terms of evaluating the performance of the developed models, besides obtaining the accuracy, which represents the number of samples classified correctly over a total

number of samples, it was also calculated the precision, recall and F_1 score of each model. Precision is the probability that an object is relevant given that it is returned by the system and recall can be defined as the probability that a relevant object is returned. As for the F_1 score, this metric is the combination of the results for precision and recall [51], [52]. The results for the accuracy, precision, recall and F_1 score of each model are in Table 19.

Table 19. Performance evaluation of each model developed.

Dataset and method used to categorise principal diagnosis	Accuracy	Precision	Recall	F_1 score
2012-2016 – ICD-9-CM main categories	0.9958	0.9972	0.9964	0.9982
2012-2016 – ICD-9-CM detailed categories	0.9975	0.9991	0.9985	0.9981
2012-2016 – CCS single-level	0.9905	0.9906	0.9999	0.9946
2012-2016 – CCS level 2	0.9870	0.9900	0.9816	0.9856
2017 – ICD-10-CM main categories	0.9821	0.9814	0.9907	0.9855
2017 – ICD-10-CM detailed categories	0.9825	0.9817	0.9770	0.9789
2017 – CCSR	0.9955	0.9966	0.9980	0.9972

CHAPTER 5 – DISCUSSION

5. Discussion

It is important to understand the main factors that lead patients to be placed and discharged for long-term care over time. In this thesis, clustering techniques and decision tree methods were used with the aim to identify and classify clusters and distinguish factors associated with LTCD.

Regarding the demographic characterization (Table 2), both databases showed similar results. Through these results, it is possible to notice an increase in the number of episodes with increasing age and a slight decrease in ages over 90 years, showing a mean age of 74 years. On the other hand, in terms of sex distribution, there are no significant differences. The sex disparity in some cases can be explained as women tend to live longer [29].

The most common principal diagnosis were “Cerebrovascular diseases”, “Pneumonia and influenza” and “Other diseases of the urinary system”. Since the population in study is mostly elderly, these diseases may be related to bedridden, demented patients with infectious complications which lead to an increase of vulnerability of these individuals and consequently, a higher level of dependency and need of hospitalization and rehabilitation. Cerebrovascular diseases such as stroke, pneumonia and influenza and urinary system diseases can lead to physical, cognitive and psychological residual disabilities constituting one of the most common reasons for hospitalization, institutionalization and need of LTC [53]–[60].

One of the most common comorbidities was hypertension. This result can be explained by the high prevalence of hypertension in the Portuguese adult population. Studies like PAP and PHYSA showed values around 42% of hypertension prevalence in Portugal. Also, hypertension is a risk factor for cerebrovascular diseases, which confirms that these diseases are among the most common principal diagnosis [61]–[64].

In fact, these factors mentioned above have been previously associated with LTCD [30]–[32], [65].

About the mode of admission, over 90% of the episodes were done through emergency services. 10% were programmed episodes like programmed surgeries, acute illness, complication or evolution of a chronic disease or deterioration of the health status needing medical and social management [11].

Concerning the methods used to detect and characterize clusters, four different approaches were applied to categorise principal diagnosis of 2012-2016 dataset and three to categorise principal diagnosis of 2017 (Figure 18).

High heterogeneity of clusters considering the main outcome variable, i.e. proportion of inpatient episodes discharged to LTC, was found. In fact, the cluster with the highest proportion ranged between 3.6% and 60.8% and between 0.7% and 5.3% for the different datasets.

However, there is no correlation between these values, i.e. clusters with the highest proportion found, and the number of categories of principal diagnosis used. In fact, while the method with the highest proportion found for the first dataset was the "ICD-9-CM detailed", the "ICD-10-CM categories" overcome the "detailed" option in the second dataset. While a conclusion is difficult to drawn from these results in terms of the hierarchical level to consider in upcoming research, it is clear that the "quality" (type of method, i.e. categorisation) overcomes "quantity" (number of categories) in identifying specific high-risk clusters.

For the 2012 to 2016 dataset, sex, being transferred from other hospital and mode of admission showed to be not relevant, once there were clusters with males and females, programmed and urgent admissions and there was only one cluster with higher proportion of LTCD in one method that had all episodes transferred from other hospital. Regarding the age, the most common was being older than 18 years. The principal diagnosis in these clusters within all methods applied were: delirium, dementia and amnestic and other cognitive disorders, diabetes mellitus with complications, other inflammatory condition of skin, viral infection, fracture of neck of femur (hip), poisoning by drugs medicinal and biological substances, endocrine nutritional and metabolic diseases and immunity disorders, diseases of the blood and blood forming organs, diseases of the circulatory system, injury and poisoning, supplementary classification of factors influencing health status and contact with health services, diseases of the digestive system or diseases of the genitourinary system. It should be noted that delirium, dementia and amnestic and other cognitive disorders were common principal diagnosis for both clusters with higher LTCD for CCS level 2 and CCS single level and these are neurodegenerative disorders that may lead to greater dependence and consequently greater need for hospitalization [66]. Metabolic equivalents, alcohol, liver, pulmonary, pulmonary circulation disorders, valvular, congestive heart failure, peripheral vascular disease, diabetes mellitus with complications, renal failure, and obesity were the comorbidities in the clusters with higher

LTCD within all methods applied. Renal failure, diabetes mellitus with complications and peripheral vascular disease were the most common, being present in the methods ICD-9-CM detailed categories and ICD-9-CM main categories.

Regarding the 2017 dataset, for all three methods applied to categorise principal diagnosis, the clusters with higher proportion of LTCD only had females. As for age, only one method used this variable as decision factor and the age category was between 65 and 79 years. The mode of admission was used but showed to be not relevant once there was one cluster with urgent admissions, one cluster with programmed admissions and one with both urgent and programmed admissions. Concerning the principal diagnosis, two methods used this variable to classify the clusters with higher proportion of LTCD and the principal diagnosis were acute haemorrhagic cerebrovascular disease or cerebral infarction and diseases of the genitourinary system. About the comorbidities, these were used by every method to classify the clusters with higher proportion of LTCD and they were hypertension, other neurological disorders, paralysis, pulmonary circulation disorders, valvular, arrhythmia or obesity. The most common within the three clusters were hypertension and valvular.

In some clusters with higher proportion of LTCD the model does not use some variables as a decision factor to be on these clusters. This may indicate that some variables such as age, comorbidities or mode of admission are more important to the model than others like principal diagnosis or being transferred from other hospital. To study and understand a possible explanation to these differences, the attribute usage of each method was studied. For the 2012-2016 dataset (Table 18), Renal, Paralysis, FluidsLytes, NeuroOther, Sex, Age, Arrhythmia and Anemia were the variables with more usage for every method. On the other hand, Lymphoma, BloodLoss, Rheumatic, Drugs and WeightLoss were the less used. For the 2017 dataset (Table 19), the more used variables were Sex, Arrhythmia, Paralysis, Liver, CHF, Age, FLuidsLytes, HTN and Alcohol. For this dataset, the variables less used by the model were Lymphoma, HIV, Psychoses, Rheumatic and BloodLoss.

About the evaluation of the model developed, analysing Table 20, it is possible to note that all methods showed values of accuracy, precision, recall and F_1 score higher than 0.900. For the 2012-2016 dataset, the method using ICD-9 detailed categories to categorise principal diagnosis attained the higher accuracy and precision values. About the recall, the method with higher value was the one using CCS single-level categories. Furthermore, the method using ICD-9-CM main categories had the higher F_1 score. On

the other hand, for the 2017 dataset, the method using CCSR to categorise principal diagnosis was the method that revealed the higher value of all four performance evaluation metrics.

As far as it is known, there are no studies in the literature comparing the results of clustering applied to different approaches to categorise principal diagnoses. Nevertheless, there are previous studies using cluster analysis to identify high-risk groups. For example, a study was developed with the aim to investigate patient multimorbidity and complexity of patients with more than 50 years [67]. Similarly, Pineda-Moncusi and his colleagues presented an unsupervised clustering technique in order to detect meaningful patterns of comorbidities in osteoarthritis patients [68].

In addition, in the literature there are some previous works where the authors applied clustering to data recorded according to ICD. For instance, with the aim to detect multimorbidity patterns, Wartelle and his colleagues proposed a hierarchical agglomerative clustering algorithm to discover diagnosis relationships from 151 recorded blocks of ICD-10 diagnoses [69]. Moreover, the same authors applied this method with the aim to analyse the impact of opening new on-demand care services based on variations in patient flow at a large hospital emergency department and the data was recorded according to the 2750 complete ICD-10 diagnostic codes [70]. Additionally, another study was developed to analyse connections between diseases based on their co-occurrences to help decision-makers in the organization of health care services. The authors used cluster analysis and for the diagnoses, they used the subgroup of ICD-10 classification (eg, I20-I25) and filtered out the ones that indicated symptoms and external causes, obtaining 205 disease groups [71]. Therefore, the methodology applied in this study proved to be useful to identify high-risk groups of patients that may have more probability of being discharged to LTC services, such as Cluster 3 of the method ICD-9-CM detailed.

CHAPTER 6 – CONCLUSIONS

6. Conclusions

With the constant increase of life expectancy, there is a grow in the number of people needing LTC. Therefore, studies analysing and determining risk factors for being placed and discharged for LTC over time are important to help prepare health care systems to solve this urgent public health problem. The aim of this thesis is to identify and classify clusters associated with LTCD which can be helpful to know patient profile that may have more probability of being to need LTC services.

Initially, a descriptive data analysis was done to get to know the sociodemographic and clinical characteristics of the inpatients analysed in this study. With this analysis it was possible to state that there was an increase in the number of episodes with the increase of the age and a slight decrease in ages over 90 years with the mean being 74 years. "Cerebrovascular diseases", "Pneumonia and influenza" and "Other diseases of the urinary system" were the most frequent principal diagnosis and hypertension was the most common comorbidity. In addition, 90% of the cases were admitted through emergency services.

Then, hierarchical clustering and C5.0 decision tree algorithm allowed to identify and characterize clusters. This showed that there are clusters with higher proportion of LTCD within each approach and there is a great variability and heterogeneity in these results depending only on the method used to categorise principal diagnosis.

Using different approaches to categorise principal diagnosis led to heterogeneous results for the identification of high-risk inpatients that are discharged to LTC. The "quality" of the principal diagnosis categorisation overcomes the "quantity" (i.e. number of categories). Nevertheless, clustering methods showed to be good options to identify high-risk groups in this case applicable to hospital discharges to LTC, having identified a group with 60.8% of episodes being discharged to LTC.

6.1. Limitations

This study has some limitations, such as those related to the quality of secondary coded data. Some episodes can possibly have errors associated to the coding of the health records, such as: miscoding/misclassification, misreporting of episodes, limited understanding of medical terminology by coders, coder experience, lack of specificity, legibility and clarity in the information registered, variations in the description of

diagnosis by clinicians, incomplete, unclear and non-specific documentation, use of synonyms and abbreviations among other problems in health records [72]–[74].

Furthermore, for the analysis only statistically valid data of inpatient episodes from 19 mainland public hospitals from Portugal which had ULDM as discharge destination were considered. It should be noted that ULDM was used as a proxy to LTC but there are another units and health care facilities providing LTC services, such as nursing homes [2].

In addition, the approach used in the selection of the optimal number of clusters can be subjective, so there may be errors associated with these values [48].

6.2. Future perspectives

In the future, it could also be useful to conduct a study with more data, including data from private hospitals, nursing homes and other units providing LTC services.

Furthermore, it would be interesting to understand the importance of the selection of the method to categorise principal diagnostics, why this model gives different results to these different approaches and which approach is more suitable to this dataset and in this context. Future analyses can also be conducted with the objective to understand if it can be applied in other studies and situations with different datasets.

Additional studies may be carried out to help predict if one patient who enters the hospital with certain characteristics, diagnosis and comorbidities may need LTC services in the future, which can be beneficial for hospital management.

7. References

[1] H. Lopes, C. Mateus, and N. Rosati, "Identifying the long-term care beneficiaries: differences between risk factors of nursing homes and community-based services admissions," *Aging Clin Exp Res*, vol. 32, no. 10, pp. 2099–2110, Oct. 2020, doi: 10.1007/s40520-019-01418-w.

[2] F. Colombo, A. Llena-Nozal, J. Mercier, and F. Tjadens, *Help Wanted? Providing and Paying for Long-Term Care*. Paris: OECD Publishing, 2011. doi: 10.1787/9789264097759-en.

[3] S. Silva, "As recentes metamorfoses da saúde na região Norte," Porto, 2008.

[4] N. Faria, "Critérios e Prioridades para a gestão financeira no Serviço Nacional da Saúde," Lisboa, Mar. 2019.

[5] L. Min and X. Huilan, "Comparative analysis of long-term care quality for older adults in China and Western countries," *Journal of International Medical Research*, vol. 48, no. 2, 2019, doi: 10.1177/0300060519865631.

[6] Ministry of Health, *Decree Law n.º 101/2006, Series I-A*. 2006, pp. 3856–3865.

[7] World Health Organization, *Lessons for Long-Term Care Policy The Cross-Cluster Initiative on Long-Term Care Noncommunicable Diseases and Mental Health Cluster World Health Organization and The WHO Collaborating Centre for Research on Health of the Elderly JDC-Brookdale Institute*. 2002.

[8] World Health Organization, *Home-based long-term care: report of a WHO Study Group*, vol. 898. Geneva: World Health Organization, 2000.

[9] ACSS-DRS, "Monitoring the Portuguese National Network for Long-term Integrated Care – 2019 (Monitorização da Rede Nacional de Cuidados Continuados Integrados - 2019) (in Portuguese)," Lisboa, 2020.

[10] S. Coutinho, "Impacto do atraso da admissão na RNCCI, dos doentes referenciados em 2016 pelos Serviços do Pólo HUC do CHUC, E.P.E.," Coimbra, Jul. 2017.

[11] J. K. Harrison *et al.*, "Predicting discharge to institutional long-term care following acute hospitalisation: a systematic review and meta-analysis," *Age Ageing*, vol. 46, no. 4, pp. 547–558, Jul. 2017, doi: 10.1093/AGEING/AFX047.

[12] Unidade de Missão para os Cuidados Continuados Integrados, "Discharge planning and management manual (Manual de planeamento e gestão de altas) (in Portuguese).," 2008. https://www.arsnorte.min-saude.pt/wp-content/uploads/sites/3/2018/05/Manual_Planeamento_Gestao_Altas.pdf (accessed Jun. 23, 2022).

[13] R. Nardi *et al.*, "Difficult hospital discharges in internal medicine wards," *Intern Emerg Med*, vol. 2, no. 2, pp. 95–99, Jun. 2007, doi: 10.1007/S11739-007-0029-7.

[14] A. H. Marshall, S. I. McClean, and P. H. Millard, "Addressing Bed Costs for the Elderly: A New Methodology for Modelling Patient Outcomes and Length of Stay," *Health Care Management Science 2004 7:1*, vol. 7, no. 1, pp. 27–33, Feb. 2004, doi: 10.1023/B:HCMS.0000005395.77308.D1.

[15] P. J. Gertler, "Medicaid and the Cost of Improving Access to Nursing Home Care," *Rev Econ Stat*, vol. 74, no. 2, pp. 338–45, May 1992, doi: 10.2307/2109668.

[16] R. F. Oliveira, "Impacto da infeção do local cirúrgico nas readmissões hospitalares de doentes ortopédicos ," Lisboa, 2018.

[17] G. Scripcaru, "Eventos adversos a medicamentos em contexto de internamento hospitalar em Portugal continental," Lisboa, 2018.

[18] D. Pimenta, "Data mining na deteção de outliers na codificação clínica: estudo em internamentos com pneumonia," Mestrado em Informática Médica, Faculdades de Medicina e de Ciências da Universidade do Porto, 2019.

[19] J. V. Santos, R. Novo, J. Souza, F. Lopes, and A. Freitas, "Transition from ICD-9-CM to ICD-10-CM/PCS in Portugal: An heterogeneous implementation with potential data implications," *Health Information Management Journal*. SAGE Publications Inc., 2021. doi: 10.1177/18333583211027241.

[20] V. Pires, "A Codificação Clínica e os problemas associados à qualidade dos dados: perspetiva dos codificadores," Mestrado em Informática Médica, Faculdades de Medicina e de Ciências da Universidade do Porto, Sep. 2018.

[21] ACSS, "Grupos de Diagnósticos Homogéneos," 2022. https://www2.acss.min-saude.pt/Default.aspx?TabId=460&language=pt-PT (accessed Oct. 29, 2022).

[22] Healthcare Cost and Utilization Project (HCUP), "Clinical Classifications Software (CCS) for ICD-9-CM," *Agency for Healthcare Research and Quality*,

Mar. 2017. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (accessed Aug. 01, 2022).

[23] Healthcare Cost and Utilization Project (HCUP), "Clinical Classifications Software Refined (CCSR) for ICD-10-PCS Procedures," *Agency for Healthcare Research and Quality*, Apr. 2022. https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/prccsr.jsp (accessed Aug. 01, 2022).

[24] W. Q. Wei *et al.*, "Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record," *PLoS One*, vol. 12, no. 7, Jul. 2017, doi: 10.1371/JOURNAL.PONE.0175508.

[25] M. Salsabili, S. Kiogou, and T. J. Adam, "The Evaluation of Clinical Classifications Software Using the National Inpatient Sample Database," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 542, 2020, Accessed: Aug. 01, 2022. [Online]. Available: /pmc/articles/PMC7233079/

[26] S. Wang, M. E. Elkin, and X. Zhu, "Imbalanced learning for hospital readmission prediction using national readmission database," *Proceedings - 11th IEEE International Conference on Knowledge Graph, ICKG 2020*, pp. 116–122, Aug. 2020, doi: 10.1109/ICBK50248.2020.00026.

[27] K. Steinbeisser, E. Grill, R. Holle, A. Peters, and H. Seidl, "Determinants for utilization and transitions of long-term care in adults 65+ in Germany: results from the longitudinal KORA-Age study," *BMC Geriatr*, vol. 18, no. 1, Jul. 2018, doi: 10.1186/s12877-018-0860-x.

[28] M. Kuzuya, S. Izawa, H. Enoki, and J. Hasegawa, "Day-care service use is a risk factor for long-term care placement in community-dwelling dependent elderly," *Geriatr Gerontol Int*, vol. 12, no. 2, pp. 322–329, Apr. 2012, doi: 10.1111/j.1447-0594.2011.00766.x.

[29] C. Y. Wu, H. Y. Hu, N. Huang, Y. T. Fang, Y. J. Chou, and C. P. Li, "Determinants of long-term care services among the elderly: A population-based study in Taiwan," *PLoS One*, vol. 9, no. 2, Feb. 2014, doi: 10.1371/journal.pone.0089213.

[30] J. K. Burton *et al.*, "Predicting Discharge to Institutional Long-Term Care After Stroke: A Systematic Review and Metaanalysis," *J Am Geriatr Soc*, vol. 66, no. 1, pp. 161–169, Jan. 2018, doi: 10.1111/jgs.15101.

[31] Y. C. Chen, W. T. Chang, C. Y. Huang, P. L. Tseng, and C. H. Lee, "Factors influencing patients using long-term care service of discharge planning by andersen behavioral model: A hospital-based cross-sectional study in eastern Taiwan," *Int J Environ Res Public Health*, vol. 18, no. 6, pp. 1–10, Mar. 2021, doi: 10.3390/ijerph18062949.

[32] A. Momose *et al.*, "Factors associated with long-term care certification in older adults: a cross-sectional study based on a nationally representative survey in Japan," *BMC Geriatr*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12877-021-02308-5.

[33] M. Macedo, "Serviço Social Hospitalar e Cuidados Continuados: Articulação entre a Saúde e o Apoio Social Informal," Coimbra, Jul. 2020.

[34] U. Brandes and T. Erlebach, *Network Analysis - Methodological Foundations*, vol. 3418 LNCS. Springer Verlag, 2005. doi: 10.1007/978-3-540-31955-9_1.

[35] P. Brugnaro *et al.*, "Clustering and risk factors of methicillin-resistant staphylococcus aureus carriage in two italian long-term care facilities," *Infection*, vol. 37, no. 3, pp. 216–221, Jun. 2009, doi: 10.1007/s15010-008-8165-1.

[36] L. van Malderen, T. Mets, P. de Vriendt, and E. Gorus, "The Active Ageing-concept translated to the residential long-term care," *Quality of Life Research*, vol. 22, no. 5, pp. 929–937, Jun. 2013, doi: 10.1007/s11136-012-0216-5.

[37] M. C. Hornbrook, "Hospital case mix: its definition, measurement and use: Part I. The conceptual framework," *Med Care Rev*, vol. 39, no. 1, pp. 1–43, 1982, doi: 10.1177/107755878203900101.

[38] A. Igarashi, N. Yamamoto-Mitani, K. Morita, H. Matsui, C. K. Y. Lai, and H. Yasunaga, "Classification of long-term care wards and their functional characteristics: Analysis of national hospital data in Japan," *BMC Health Serv Res*, vol. 18, no. 1, Aug. 2018, doi: 10.1186/s12913-018-3468-0.

[39] R. Woodhouse, J. K. Burton, N. Rana, Y. L. Pang, J. E. Lister, and N. Siddiqi, "Interventions for preventing delirium in older people in institutional long-term care," *Cochrane Database of Systematic Reviews*, vol. 2019, no. 4. John Wiley and Sons Ltd, Apr. 23, 2019. doi: 10.1002/14651858.CD009537.pub3.

[40] J. Chrusciel *et al.*, "Preparing the transition into long-term care institutions: a cluster analysis of the representations of the place of residence in European elders," *Res Sq*, 2019, doi: 10.21203/rs.2.16670/v1.

[41] K. S. Boockvar, K. M. Judon, J. P. Eimicke, J. A. Teresi, and S. K. Inouye, "Hospital Elder Life Program in Long-Term Care (HELP-LTC): A Cluster Randomized Controlled Trial," *J Am Geriatr Soc*, vol. 68, no. 10, pp. 2329–2335, Oct. 2020, doi: 10.1111/jgs.16695.

[42] D. Martinho *et al.*, "A Hybrid Model to Classify Patients with Chronic Obstructive Respiratory Diseases," *J Med Syst*, vol. 45, no. 3, Mar. 2021, doi: 10.1007/S10916-020-01704-5.

[43] A. F. Lacerda, G. Oliveira, C. Cancelinha, and S. Lopes, "Hospital inpatient use in mainland Portugal by children with complex chronic conditions (2011 – 2015)," *Revista científica da Ordem dos Médicos - Acta Médica Portuguesa* , vol. 32, no. 7–8, pp. 488–498, Jul. 2019, doi: 10.20344/amp.10437.

[44] M. Fialho, "Hospitalizations due to Diabetes in Portugal: a time series analysis," Lisboa, 2020.

[45] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Med Care*, vol. 36, no. 1, pp. 8–27, 1998, doi: 10.1097/00005650-199801000-00004.

[46] A. Freitas, I. Lema, and A. D. da Costa-Pereira, "Comorbidity coding trends in hospital administrative databases," *Advances in Intelligent Systems and Computing*, vol. 445, pp. 609–617, 2016, doi: 10.1007/978-3-319-31307-8_63/COVER.

[47] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis 5th Edition Cluster Analysis 5th Edition WILEY SERIES IN PROBABILITY AND STATISTICS Cluster Analysis 5th Edition*, 5th ed. 2011.

[48] M. Roux, "A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms," *Journal of Classification* , vol. 35, no. 2, pp. 345–366, Aug. 2018, doi: 10.1007/S00357-018-9259-9.

[49] J. Basak and R. Krishnapuram, "Interpretable hierarchical clustering by constructing an unsupervised decision tree," *IEEE Trans Knowl Data Eng*, vol. 17, no. 1, pp. 121–132, Jan. 2005, doi: 10.1109/TKDE.2005.11.

[50] P. Sujatha and K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*, Nov. 2020, doi: 10.1109/INOCON50539.2020.9298354.

[51] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," *Lecture Notes in Computer Science*, vol. 3408, pp. 345–359, 2005, doi: 10.1007/978-3-540-31865-1_25/COVER/.

[52] W. Chen *et al.*, "Comprehensive geriatric functional analysis of elderly populations in four categories of the long-term care insurance system in a rural, depopulated and aging town in Japan," *Geriatr Gerontol Int*, vol. 13, no. 1, pp. 63–69, Jan. 2013, doi: 10.1111/J.1447-0594.2012.00859.X.

[53] S. K. Lui and M. H. Nguyen, "Elderly Stroke Rehabilitation: Overcoming the Complications and Its Associated Challenges," *Curr Gerontol Geriatr Res*, vol. 2018, 2018, doi: 10.1155/2018/9853837.

[54] P. Winnige, R. Vysoky, F. Dosbaba, and L. Batalik, "Cardiac rehabilitation and its essential role in the secondary prevention of cardiovascular diseases," *World J Clin Cases*, vol. 9, no. 8, p. 1761, Mar. 2021, doi: 10.12998/WJCC.V9.I8.1761.

[55] E. Westerlind, D. Hörsell, and H. C. Persson, "Different predictors after stroke depending on functional dependency at discharge: a 5-year follow up study," *BMC Neurol*, vol. 20, no. 1, Jul. 2020, doi: 10.1186/S12883-020-01840-Y.

[56] Y. Iwamoto *et al.*, "Development and Validation of Machine Learning-Based Prediction for Dependence in the Activities of Daily Living after Stroke Inpatient Rehabilitation: A Decision-Tree Analysis," *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 12, p. 105332, Dec. 2020, doi: 10.1016/J.JSTROKECEREBROVASDIS.2020.105332.

[57] P. Moyo *et al.*, "Risk factors for pneumonia and influenza hospitalizations in long-term care facility residents: A retrospective cohort study," *BMC Geriatr*, vol. 20, no. 1, pp. 1–13, Feb. 2020, doi: 10.1186/S12877-020-1457-8/TABLES/3.

[58] E. Bosco *et al.*, “Long-term Care Facility Variation in the Incidence of Pneumonia and Influenza,” *Open Forum Infect Dis*, vol. 6, no. 6, Jun. 2019, doi: 10.1093/OFID/OFZ230.

[59] L. Genao and G. T. Buhr, “Urinary Tract Infections in Older Adults Residing in Long-Term Care Facilities,” *Ann Longterm Care*, vol. 20, no. 4, p. 33, Apr. 2012, Accessed: Aug. 05, 2022. [Online]. Available: /pmc/articles/PMC3573848/

[60] A. P. Serafim, A. L. Martins-Ferreira, M. P. Serafim, G. Oliveira, E. Pedro-Rocheta, and N. Pires, “Prevalência da hipertensão arterial na população portuguesa em contexto de férias e abordagem multivariada dos fatores de risco através do método HJ-Biplot.,” *Revista Portuguesa de Medicina Geral e Familiar*, vol. 35, no. 6, pp. 450–64, Dec. 2019, doi: 10.32385/RPMGF.V35I6.12319.

[61] D. Costa and R. Peixoto Lima, “Custo-efetividade da monitorização ambulatória da pressão arterial na abordagem da hipertensão arterial,” *Revista Portuguesa de Cardiologia*, vol. 36, no. 2, pp. 129–139, Feb. 2017, doi: 10.1016/J.REPC.2016.09.007.

[62] M. E. Macedo, M. J. Lima, A. O. Silva, P. Alcantara, V. Ramalhinho, and J. Carmona, “Prevalence, awareness, treatment and control of hypertension in Portugal: The PAP study,” *J Hypertens*, vol. 23, no. 9, pp. 1661–1666, 2005, doi: 10.1097/01.hjh.0000179908.51187.de.

[63] J. Polonia, L. Martins, F. Pinto, and J. Nazare, “Prevalence, awareness, treatment and control of hypertension and salt intake in Portugal: changes over a decade. The PHYSA study,” *J Hypertens*, vol. 32, no. 6, pp. 1211–1221, 2014, doi: 10.1097/HJH.0000000000000162.

[64] K. Lee and E. Cho, “Activities of daily living and rehabilitation needs for older adults with a stroke: A comparison of home care and nursing home care,” *Japan Journal of Nursing Science*, vol. 14, no. 2, pp. 103–111, Apr. 2017, doi: 10.1111/jjns.12139.

[65] R. E. Hales, *The American Psychiatric Publishing Textbook of Psychiatry* , 5th ed. 2008. Accessed: Oct. 04, 2022. [Online]. Available: https://books.google.pt/books?hl=pt-PT&lr=&id=2RzFWRIAsPAC&oi=fnd&pg=PA303&dq=delirium+dementia+and+amnestic+and+other+cognitive+disorders&ots=KUiPCcw_JW&sig=U_xxA9e

4JU9YIl0bPSKNMxHmbRE&redir_esc=y#v=onepage&q=delirium%20dementia%20and%20amnestic%20and%20other%20cognitive%20disorders&f=false

[66] A. Nicolet *et al.*, "Exploring Patient Multimorbidity and Complexity Using Health Insurance Claims Data: A Cluster Analysis Approach," *JMIR Med Inform*, vol. 10, no. 4, Apr. 2022, doi: 10.2196/34274.

[67] M. Pineda-Moncusi, V. Strauss, D. Robinson, D. Prieto-Alhambra, and S. Khalid, "Unsupervised Learning to Understand Patterns of Comorbidity in 633,330 Patients Diagnosed with Osteoarthritis," pp. 121–129, Mar. 2022, doi: 10.5220/0010833500003123.

[68] A. Wartelle, F. Mourad-Chehade, F. Yalaoui, J. Chrusciel, D. Laplanche, and S. Sanchez, "Clustering of a Health Dataset Using Diagnosis Co-Occurrences," *Applied Sciences 2021, Vol. 11, Page 2373*, vol. 11, no. 5, p. 2373, Mar. 2021, doi: 10.3390/APP11052373.

[69] A. Wartelle *et al.*, "Multimorbidity clustering of the emergency department patient flow: Impact analysis of new unscheduled care clinics," *PLoS One*, vol. 17, no. 1, p. e0262914, Jan. 2022, doi: 10.1371/JOURNAL.PONE.0262914.

[70] P. Fränti, S. Sieranoja, K. Wikström, and T. Laatikainen, "Clustering Diagnoses From 58 Million Patient Visits in Finland Between 2015 and 2018," *JMIR Med Inform 2022;10(5):e35422 https://medinform.jmir.org/2022/5/e35422*, vol. 10, no. 5, p. e35422, May 2022, doi: 10.2196/35422.

[71] R. Carvalho *et al.*, "Analysis of root causes of problems affecting the quality of hospital administrative data: A systematic review and Ishikawa diagram," *Int J Med Inform*, vol. 156, p. 104584, Dec. 2021, doi: 10.1016/J.IJMEDINF.2021.104584.

[72] V. Alonso *et al.*, "Problems and Barriers during the Process of Clinical Coding: a Focus Group Study of Coders' Perceptions," *Journal of Medical Systems 2020 44:3*, vol. 44, no. 3, pp. 1–8, Feb. 2020, doi: 10.1007/S10916-020-1532-X.

[73] V. Alonso *et al.*, "Health records as the basis of clinical coding: Is the quality adequate? A qualitative study of medical coders' perceptions," *Health Inf Manag*, vol. 49, no. 1, pp. 28–37, Jan. 2020, doi: 10.1177/1833358319826351.