Universidade de Aveiro

2022

Beatriz Maria Moutinho da Costa e Carvalho Ribeiro

Analisar o impacto da Multimorbilidade em outcomes de saúde, entre hospitalizações por cancro do estômago

Analysing the impact of Multimorbidity on health outcomes among stomach cancer hospitalizations

Universidade de Aveiro

2022

Beatriz Maria Moutinho da Costa e Carvalho Ribeiro

Analysing the impact of Multimorbidity on health outcomes among stomach cancer hospitalizations

Dissertação apresentada à Universidade de Aveiro no âmbito de estágio realizado no Centro de Investigação em Tecnologias e Serviços de Saúde (CINTESIS), para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Biomédica, realizada sob a orientação científica do Professor Doutor Sérgio Guilherme Aleixo de Matos, professor auxiliar do Departamento de eletrónica, telecomunicações e informática da Universidade de Aveiro (UA), e do Doutor Júlio Cesar Botelho de Souza, investigador no CINTESIS.

o júri / the jury

Professora Doutora Ana Luísa Monteiro da Silva
Professora Auxiliar em Regime Laboral, Universidade de Aveiro

vogais / examiners committee

Arguente Principal: Doutora Cláudia Camila Rodrigues Pereira Dias
Equiparada a Investigadora Auxiliar, Faculdade de Medicina da Universidade do Porto

Coorientador: Doutor Júlio Cesar Botelho de Souza
Investigador, Faculdade de Medicina da Universidade do Porto

agradecimentos

I would like to thank CINTESIS - Center for Research in Health Technologies and Services for their collaboration in recent months. I appreciate the opportunity to expand my knowledge in Medical Informatics, on a highly relevant and interesting topic.

I want to present my gratitude to my main supervisor Julio Souza for his constant availability and time spent throughout the work.

Thanks to Professor Alberto Freitas for the initial integration into the work and feedback over the months, and also to doctors João Vasco Santos and Diogo Libânio for sharing their knowledge in the field of medicine and contextualizing multimorbidity and stomach cancer.

In the end, I cannot aimless the crucial role of my family and friends, who have supported me and given me the strength to overcome difficulties.

.

palavras-chave

Multimorbilidade, cancro do estômago, comorbilidades de Charlson, padrões, outcomes, mortalidade, tempo de internamento

resumo

O envelhecimento da população tem sido um grande desafio para os sistemas de saúde, devido ao aumento de pacientes com multimorbidade, que, por sua vez, necessitam de cuidados mais complexos e visitas regulares aos serviços de saúde. O cancro de estômago é uma das neoplasias mais comuns em todo o mundo, portanto, as pesquisas têm vindo a centrar-se em comorbidades específicas e respetivos efeitos na trajetória clínica de um paciente multimórbido com essa doença. Assim, novos insights sobre diagnóstico, tratamento e vigilância do cancro serão fornecidos, a longo prazo. Entre os padrões de multimorbidade obtidos por algoritmos de Machine Learning não supervisionados, os Diabetes sem complicações, a Doença Pulmonar Obstrutiva Crónica e a Insuficiência Cardíaca Congestiva foram a combinação mais comum de comorbilidades de Charlson obtidas tanto no clustering, como no agrupamento, entre 2011 e 2015. Na visualização em rede, também foi destacada a elevada correlação de cada uma destas comorbilidades com as restantes utilizadas neste trabalho. Em relação à mortalidade, o cancro de estômago do tipo piloro/antro pilórico apresentou menor risco de morte (OR 0,88, IC 95% 0,79 – 0,98) em comparação com os tipos cárdia, fundo e corpo. Além disso, pacientes com cancro de estômago expostos a tumores metastáticos e com quatro ou mais comorbidades revelaram um risco significativamente maior de morte. A região superior do estômago apresentou taxas de mortalidade significativamente maiores na presença de tumor sólido metastático (OR 2,95, IC 95% 2,41 – 3,62) e na presença de quatro ou mais comorbidades (OR 1,62, IC 95% 0,95 – 2,77), enquanto para a região superior do estômago, esses fatores também desempenharam um papel importante na mortalidade [(OR 2,64, IC 95% 2,18 – 3,20) e (OR 3,63, IC 95% 2,14 – 6,16), respetivamente]. Em particular, a região superior do estômago apresentou risco significativamente maior de morte para pacientes com mais de 75 anos (OR 1,69, IC 95% 1,21 – 2,35). Por outro lado, as neoplasias malignas do piloro/antro pilórico do estômago (IRR 1,07, IC 95% 1,06 – 1,08) apresentaram-se intimamente relacionadas a internações mais longas em comparação com as neoplasias malignas do cárdia/fundo/corpo. Outros fatores que contribuíram para o aumento do tempo de internamento incluíram a presença de três comorbidades (IRR 1,38, IC 95% 1,32 – 1,44), duas comorbidades (IRR 1,36, IC 95% 1,32 – 1,40) e envelhecimento (IRR 1,11, 95% IC 1,07 – 1,16), considerando internações para região cancro da região superior do estômago. A principal diferença entre as regiões do estômago é que a região inferior apresentou internações mais longas em pacientes com a presença de três (IRR 1,25, IC 95% 1,20 – 1,30) e quatro ou mais comorbidades (IRR 1,25, IC 95% 1,17 – 1,34)). Resumidamente, nos próximos anos, os cuidadores de saúde em Portugal devem focar-se na presença de duas ou mais comorbilidades, em doentes com cancro do estômago, especificamente no sexo masculino e em estágios avançadas do cancro do estômago (metástase tumoral), uma vez que estes fatores aumentam as probabilidades de morte. Explicitamente, doenças como Diabetes Não Complicado, Insuficiência Cardíaca Congestiva e Doença Pulmonar Obstrutiva Crônica requerem especial atenção na presença de neoplasia maligna de estômago. Além disso, é notável a necessidade de reforçar a gestão de recursos hospitalares para pacientes multimórbidos com cancro de estômago do tipo pilórico/antro, pilórico, uma vez que está associado ao aumento do tempo de internamento (proxy do uso de recursos), bem como um melhor diagnóstico, prevenção e tratamento de pacientes multimórbidos com neoplasma maligno da cardia do estômago, que está associado ao aumento da mortalidade.

keywords

Multimorbidity, Stomach cancer, Charlson' comorbidities, patterns, outcomes, mortality, length of stay

abstract

The aging of the population has been a major challenge for health systems, due to the increase in patients with multimorbidity, who require more complex treatments and extensive use of health services. Stomach cancer is one of the most common neoplasms worldwide, so research has focused on specific comorbidities and their effects on the clinical trajectory of a patient with this disease. Hence, new insights into optimal diagnosis, treatment, and surveillance of cancer patients with comorbid disease will be generated in long-term period. Among patterns of multimorbidity obtained by unsupervised Machine Learning algorithms, uncomplicated diabetes, Chronic Obstructive Pulmonary Disease, and Congestive Heart Failure were the most common combination of Charlson' comorbidities obtained in both clustering and association rules, between 2011 and 2015. For network visualization, the high correlation of each of these comorbidities with the others used in this work was also highlighted. Regarding death outcomes, pylorus/pyloric antrum stomach cancer showed a lower risk of death (OR 0.88, 95% CI 0.79 – 0.98) in comparison to malignant neoplasms of cardia, fundus and body of stomach. Also, stomach cancer patients exposed to metastatic tumors and four or more comorbidities revealed a significantly higher risk of death. The upper region of the stomach presented significantly higher odds of death rate under the presence of metastatic solid tumor (OR 2.95, 95% CI 2.41 – 3.62) and the presence of four or more comorbidities (OR 1.62, 95% CI 0.95 – 2.77), whereas for the upper region of the stomach, these factors also a played a major role in the odds of death [(OR 2.64, 95% CI 2.18 – 3.20) and (OR 3.63, 95% CI 2.14 – 6.16), respectively]. In particular, the upper region of the stomach also presented a significantly higher risk of death for people over 75 years (OR 1.69, 95% CI 1.21 – 2.35). On the other hand, cancers affecting pylorus/pyloric antrum (IRR 1.07, 95% CI 1.06 – 1.08) were related to longer hospitalizations in comparison with cardia/fundus/body. Other factors contributing with the increase of length of stay included the presence of three comorbidities (IRR 1.38, 95% CI 1.32 – 1.44), two comorbidities (IRR 1.36, 95% CI 1.32 – 1.40), and aging (IRR 1.11, 95% CI 1.07 – 1.16), considering hospitalizations in which cancer affected the upper stomach region. The main difference between regions of the stomach is that the lower region presented longer hospitalizations in patients with the presence of three (IRR 1.25, 95% CI 1.20 – 1.30), and four or more comorbidities (IRR 1.25, 95% CI 1.17 – 1.34). Briefly, in the next years, Portugal' health care stakeholders need to focus on presence of two or more comorbidities in stomach cancer patients, specifically on male sex and in advanced stage of cancer (tumor metastasis) as these factors increases the odds of death. Explicitly, diseases such diabetes, CHF and COPD require special attention in the presence of malignant neoplasm of stomach. Also, is notable a demand for a reinforced hospital' resources management for multimorbid Pyloric/antrum stomach cancer patients, which is associated with increased length of stay (a proxy of resource use), and for a better diagnosis, prevention and treatment for multimorbid Cardia stomach cancer, which is associated with increased mortality.

Contents

List of tables

List of figures

1. Introduction

1.1. Contextualization

Currently, health services in most countries recognize that a patient with more than one health problem, the so-called multimorbid patients, need personalized treatment, in which chronic diseases should not be treated in isolation. With the improvement in life expectancy, the population tends to accumulate several diseases in recent decades, so the prevalence of multimorbidity, worldwide, has increased dramatically. Cancer patients have a very poor prognosis, especially in advanced stages. In this sense, the presence of multimorbidity in this condition urgently requires well-oriented health planning, which will refine hospital management and policymakers to provide services that can effectively reduce the cancer burden in the coming years.

In this sense, the present work suggests an approach to the analysis of multimorbidity patterns in patients with stomach cancer, the fifth most common cancer worldwide, and the consequent relationship and impact of multimorbidity on health outcomes in those patients.

Furthermore, in the context of my Biomedical Engineering Master, this dissertation allowed deepening my knowledge of data mining, such as machine learning, learned throughout my academic path, and to apply this knowledge in healthcare situations to help professionals in decision-making and provide guidance for further investigations on the burden of stomach cancer in clinical practice.

1.2. Cancer of the stomach

Functional Anatomy

The stomach constitutes an important organ and the most distensible portion of the digestive system, in which the early phases of food digestion begin, by acid enzyme secretion. After being initiated, the stomach will deliver ingested nutrients via rhythmic motion to the small intestine[1],[2].

Anatomically, it is divided into five regions, the cardia, fundus, corpus, antrum, and pylorus, each of which with distinctive structures that promote specific functions **(Figure 1)**[3]. Across them, a protective layer of columnar epithelial cells covers the lining. The stomach wall is constituted by 4 different layers, in descending order of profundity: mucosa, submucosa, muscularis propria and serosa. The mucosa on the lesser curve and in transition zones (antrum-body, cardia-body) is commonly thinner than on the greater curve[1],[2].

The cardia is the junction between the lower oesophagus and the stomach and is where the food first enters. The fundus is a dome on the left, which results from the extension of the cardia. The body is the largest portion of the stomach, where the food is mixed, and extends from the fundus to the incisura angularis. The gastric antrum serves to grind food into smaller particles that are sieved into the duodenum. The pylorus connects the stomach directly to the duodenum, which forms the first segment of the small intestine. Also, pylorus contains pyloric sphincter that prevents the reflux of duodenal content back into the stomach transporting small portions of acidic chyme to pass into it[1],[2].

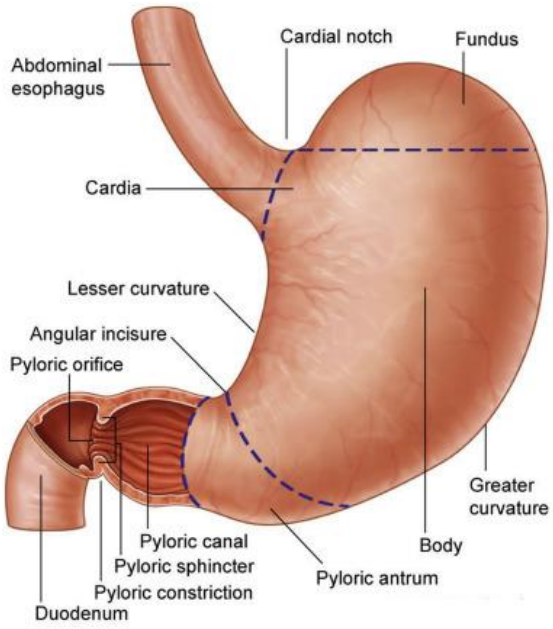

Figure 1 – Representation of different stomach structures. It contains four parts: cardia, Fundus., Body and Pyloric, two ends: Cardiac and pyloric and two curvatures: Lesser and greater. Image extracted from[3]

Etiology

Gastric carcinogenesis involves a cascade of occurrences (Correa gastric carcinogenesis cascade) that can be modulated by various factors. The first occurrence is normally potentiated by *Helicobacter pylori* infection, a *Gram-negative bacteria* that colonises gastric mucosa and cause chronic gastritis virtually in all infected individuals and peptic ulceration in a subset of patients. Specifically, it directs the *cag gene*, positive to inflammation in the stomach, and *VacA* cytotoxin generates epithelial cell damage[4]. Other distinctive risk factors for gastric cancer are diet, smoking, alcohol intake, environmental agents, and genetic background. Related to the diet, high sodium intake acts directly on the stomach lining, destroying the mucosal barrier and causing gastritis, while lack of intake of fresh fruits and vegetables reduces the presence of antioxidants and, consequently, the protection against oxidative damage. Tobacco smoke carcinogens affect gastric cancer risk directly through contact with the stomach mucosa or indirectly through the blood flow[5]. Additionally, alcohol consumption results in the formation of acetaldehyde, a carcinogen that can directly damage the gastric mucosa. Related to the environment, some chemical agents or radiation exposure have been linked to the development of stomach cancer. Finally, hereditary stomach cancer takes only 1–3% of all this cancer, but having access to detailed family history and tumours' histological classification to identify stomach cancer predisposition from birth remains an important matter regarding prevention [4],[6],[7].

The long-term consequence of *H. pylori* chronic gastritis is the development of precancerous conditions (gastric atrophy and/or intestinal metaplasia) in a subset of infected individuals. Indeed, most cases of gastric adenocarcinoma of intestinal-type (the most common form) develop on gastric mucosa with multifocal atrophic gastritis (loss of gastric glands) and /or extensive intestinal metaplasia (substitution of gastric epithelium by intestinal glands), suggesting that both are premalignant conditions of the stomach. Atrophic mucosal changes usually develops first in the antrum and progress to the corpus along the lesser curvature[8].

The different types of stomach cancer vary on epidemiologic relationships, associated risk factors, and prognosis. The majority of stomach cancer cases (90–95%) are gastric carcinomas (malignant epithelial

neoplasms), and adenocarcinomas are the overwhelming majority of them[9]. The remaining 5% are composed of lymphoma, carcinoid, leiomyosarcoma, and squamous cell carcinoma[10].

Gastric adenocarcinomas are classified anatomically as proximal (cardia) and distal (non-cardia), both with different molecular pathogenesis, biologic behaviour, and clinical outcomes. The proximity of cardia to esophagogastric junction (EGJ) implies some shared etiologies with distal oesophageal adenocarcinomas[9]. In this sense, the main cause for cardia adenocarcinoma is gastroesophageal reflux disease and obesity, while 90% of non-cardia cancers are attributable to *Helicobacter pylori* (H. pylori) infection[11].

Epidemiology and Pathology

Despite a reduction in mortality during the last years, stomach cancer is the third cause of cancer death worldwide and the fifth most common cancer[11]. Stomach cancer resulted in 1.3 million (1.2–1.4 million) incident cases, 9.5 hundred thousand (8.7–10.4 hundred thousand) deaths, and 22.2 million (20.3–24.1 million) DALYs (Disability-adjusted life years) in 2019[12]. Also, global patterns of gastric cancer in 2020 revealed about 1.1 million new cases and 770,000 deaths[11]. In order to forecast the future progress of this disease, the annual burden of gastric cancer suggested an increase of approximately 1.8 million new cases and 1.3 million deaths by 2040[11].

According to several studies, stomach cancer mostly affects older people, with an average age of 68 years. Although it is considered a rare disease in young individuals, recent reports have shown that 2-15% of gastric cancer cases are diagnosed in individuals 50 years old or less, particularly for non-cardia tumours in wealthier populations[11]. A more detailed analysis by DALYs publicised the highest percentages in the 65–69 year age group for smoking and the 50–54 year age group for a high-sodium diet, and they were the lowest in the 30–34 year age group for smoking and in the 95 and over group for a high-sodium diet[11].

Worldwide, the distribution is uneven, considering that more than 70% of the cases occur in developing countries, and half of the worldwide total cases occurs in eastern Asia. Eastern Asian and Eastern European regions showed the highest incidence rates in both males and females and globally about two thirds of all cases occur in men [10]. Other reports suggest the highest incidence among all cancers for males in which, for gastric cardia carcinomas, men are affected five times more than women [6],[12]. Moreover, North America and most countries in Africa and Southeast Asia are problematic[13]. The highest survival rate in Europe belongs to Iceland, which reports a 42% 5-year survival rate among women [7]. In contrast, Portugal admits the highest gastric cancer mortality rates in Western Europe, due to the increasing prevalence of Helicobacter pylori infection[14].

In recent years, rates of distal gastric cancers have declined, coinciding with the widespread treatment of *H. pylori*, improved sanitation, refrigeration leading to less smoking and salt preservation of food, and more varied diets of fresh fruits and vegetables. Also, a prevention process for decreasing the risk of gastric cancer in infected individuals without premalignant lesions, includes H. pylori eradication [14]. Unfortunately, low-incidence countries still admit higher rates of cancers of the proximal stomach and esophagogastric junction (OGJ)[7]. However, non-cardia stomach cancer continues to be diagnosed twice as often as cardia[7]. Also, it was suggested more prevalence of cancers of the antrum and pylorus in regions of high incidence[8].

Instead, the cost associated with stomach cancer management per patient remains higher, so that a consistent and systematic analysis of the global long-term trends and patterns of stomach cancer is essential to guide policymakers for proper decision-makers [11].

Diagnosis

Carcinoma of the stomach may be detected either in an early stage or in an advanced stage, but the majority of patients with early gastric cancer are asymptomatic. Upper endoscopy is a common primary non-invasive test usually applied for screening and diagnosis[11].

Considering symptomatic patients, some signs are described such as nausea, vomiting, dysphagia, abdominal pain, difficulty swallowing, unexplained weight loss, and gastrointestinal bleeding. If detected early stages, with low risk of lymph node metastasis, early gastric cancer can be removed by endoscopic submucosal dissection with advantages of lower costs and morbidity and a lower impact in health-related quality of life[8]. In that sense, these patients are associated with favourable prognosis, with a 5-year survival rate higher than 85% [7].

Nonetheless, and since there are no established screening programs in Portugal, most gastric cancers are diagnosed in an advanced stage, with lymph-node metastases and vascular invasion. The vast majority of H pylori-infected individuals remain asymptomatic without any clinical sequelae [14],[15].

Unfortunately, lymphatic and vascular invasion, often seen in advanced cases, specifically carries a poor prognosis [13]. In fact, with 5-years survival <50% even when treated with multimodal therapy (surgery ± chemotherapy ± radiotherapy) [1],[4].

1.3. Multimorbidity

1.3.1. Definition of multimorbidity

The rising of longevity is one of the main causes of the aging population[16]. An important aspect of this scenario is the accumulation of multiple diseases in one individual, which has been a common phenomenon in clinical and community settings[17], [18]. The term comorbidity was proposed in 1970 to describe the co-occurrence of additional conditions alongside a primary or index condition[19]. For years, etiologic interventions designed for comorbidity have made it possible to assess the influence of the coexistence of one or more disorders on a primary disease of interest. In this sense, the concept of comorbidity is best applied to secondary and tertiary care[20],[21]. However, the evolution of health systems required the exploration of potentially causal associations between all coexisting conditions at once. This allows to create a map of common patterns and the groups of concomitant diseases[22].

In that manner, a new definition comes out, that considers health-related characteristics but also socioeconomic, cultural, environmental, and behaviour factors[23]. Multimorbidity was firstly defined by the World Health Organization (WHO) (2008), as the presence of multiple chronic health conditions (i.e., physical and mental) treated to consider all conditions in one individual that could affect his global health status[18]. In 2018, a new definition was presented by the Academy of Medical Science, as a standardized one to clear the classification and try to make different research findings more comparable [18]. The multimorbidity condition was presented as the existence of two or more ‘long-term’ or ‘chronic’ conditions, at the same time, which can be 1) a physical non-communicable disease of long duration; 2) a mental health condition of long duration; or 3) an infectious disease of long duration. In that manner, it constitutes a really concerning situation because of the negative impact on the person's life, such as reduced quality of life, increased frailty and inability to perform daily tasks and manage medication, the rise of hospital admissions and permanent care[19],[21],[23].

Unfortunately, there isn´t a current consensus about multimorbidity designation in the context of clinical care, epidemiological research, or health service planning [21]. Therefore, their measurements are complex as they are dependent on the population studied, outcome of interest and the context in which it is used[16]. Detailed information is presented in following section.

1.3.2. Measuring multimorbidity and functional limitation

Despite the growing interest in multimorbidity, his access constitutes a challenge for clinical management. This is explained by the lack of methodological measure standardisation, such as operational definitions, choice of study populations and data sources[18], [21].

Concretely, there is a large variation in the number of conditions included in multimorbidity measures, and which conditions were included [16]. Following [24] the most chronic conditions included in studies are diabetes, stroke, cancer, chronic obstructive pulmonary disease, hypertension, coronary heart disease, chronic kidney disease, and heart failure. However, the number can range up to 185 different diseases[25].

Furthermore, methods to collect the information differ between clinical records in hospitals or primary care units, while others collected information through self-report[19].

Regular multimorbidity data sources include clinical administrative databases, self-administered surveys, and interview-based surveys. In these databases, the chronic conditions are preferably coded, to be distinguished from functional deficits or disabilities, frailty, or other states of poor health [26], [27].

The worldwide classification and clinical coding system is the International Classification of Diseases (ICD), developed by the World Health Organization (WHO), defined as "the standard diagnostic tool for epidemiology, health management, and clinical purposes" including "the analysis of the general health situation of population groups". This classification system arranges health conditions into a hierarchical structure that can be used by many countries [27],[28].

Measures of multimorbidity broadly fall into 2 types: simple counts of diseases, in each individual, and weighted index [29]. During the years, multimorbidity was commonly assessed by counting the number of morbidities, based on patient self-report or clinician assessment, particularly, for estimating its prevalence[30],[31]. However, self-reported information does not reflect patients' experience or the effects of different combinations or severity of diseases [19],[16]. Given this inadequacy, weighted indexes evolved to assess the severity or level of impairment caused by a disease, called the burden of disease [19],[32]. They can predict relevant outcomes, such as daily functioning and quality of life, mortality risk, health service use and postoperative complications, when considered as a dependent variable in different models [33],[34].

The Charlson Comorbidity Index (CCI) is the most widely used scoring system for comorbidities used by researchers and clinicians, as a short and long-term outcome such as function, hospital length of stay and mortality rates. It has been adapted to administrative databases, such as the International Classification of Diseases, 9th (ICD-9) and 10th (ICD-10) codes, medical procedures, and medication[30]. An adaptation of CCI is Elixhauser Comorbidity Index, with additional versatility (covering acute and chronic conditions) to predict outcomes[35]. In 1968, a new measure was developed, the Cumulative Disease Rating Scale (CIRS), differing from the previous ones by assessing the severity of disability in 13 areas grouped by body systems[21]. Later, with the growing need to monitor healthcare usage and costs, Adjusted Clinical Groups (ACG) emerged as a system capable of combining data from a variety of sources and generating insights beyond mortality and quality of life[29]. A more detailed comparison is made in **Table 1.**

Further, an overview of the number of diseases included in some multimorbidity indices revealed that Diabetes mellitus is the most frequent single diagnosis listed in multimorbidity indices, followed by stroke, cardiovascular diseases, hypertension, cancer, chronic obstructive pulmonary disease (COPD), and (osteo)arthritis. Less common conditions include kidney diseases, heart failure, myocardial infarction, and depression [36].

Table 1-Comparative table of Multimorbidity indices [20], [22] ,[28], [35]

	Disease count	*Charlson Index*	*Elixhauser index*	*ACG system*	*The Cumulative Illness Rating Score (CIRS)*
Description	Derived from medical records or clinician diagnosis Self-reported disease counts based on questionnaires or interviews	19 chronic conditions weighted 1, 2, 4, or 6 based on outcome	Similar system of Charlson, but includes 30 conditions, each weighted 1 point	Stratify the population in groups by medical records with similar clinical diagnosis, age, and sex	Score system applied to each of 14 independent body systems, rated from 1 (no impairment to that organ/ system) to 5 (impairment is life threatening).
Outcome	Quality of life	Quality of life and hospital mortality within 1 year of hospitalization	Hospital mortality, length of stay and adverse events	Future morbidity and health resources' utilization	Medical burden of chronic illness
Limitations	Variety of disease' lists in different studies and a difficult assess due to limited data	Requires specific ICD coding (beyond 3 digits) for accuracy	Comorbidities limited to ICD codes recorded for index admission only	Non transparent scoring systems which often consider costs to end-users	The classification system and prognostic indicators themselves include, only, a few indices that assess the diversity of diseases

1.3.3. Prevalence and patterns of multimorbidity

Prevalence

A prevalent disease is defined as a specific disease in a subject that had been diagnosed by a doctor and was being treated at the time of the survey. To understand disparities among multimorbid populations, it requires the consistent monitoring of the populations by age, gender, race and ethnicity, geographic factors, socioeconomic status, and physical environment [37].

Once again by virtue of the lack of generalization across studies, investigating the prevalence of multimorbidity becomes, also, a difficult task. Specifically, variability encompasses the method of recruitment and sample size, data collection, sociodemographic settings, and the number of diagnoses considered in the definition of multimorbidity [32],[38].

According to a revision of 193 studies of prevalence analysis [39], a consensual prevalence value of multimorbidity was 42.4% , with high heterogeneity, by differences between studies. A wind up was a strong association between older age and a larger number of conditions with a higher prevalence of multimorbidity. For real, the global prevalence of multimorbidity is expected to increase through the 21st century, as a result of increased life expectancy and population aging [21], [40]. Thus, it is expected that the prevalence of multimorbidity is higher when the number of conditions eligible for inclusion in the definition is higher. Instead, sociodemographic factors such as sex did not reveal significant differences in prevalence, although in some studies the prevalence is greater in women.

In low- and middle-income countries, compassed by social disadvantages and fragile healthcare systems, the condition of multimorbidity is rising and impacts younger people too[41]. Families with lower socioeconomic levels and predominance of non-communicable chronic diseases produce rapid shifts in deleterious effects on health in many low-income nations[42],[41]. In parallel, T J Bolyky et *al*. reported cancers, diabetes, cardiovascular diseases and chronic respiratory illnesses as continuously rising chronic diseases in that countries [43].

Tran et *al.* [44] reveals that, across Europe, many countries have seen 40–60% of those aged 50 years or older living with multimorbidity. Another European analysis, during the period 2004–2017, found Cardiometabolic and musculoskeletal diseases were more prevalent while cancer and neurodegenerative

diseases were less prevalent. In 2021, Koné et *al.* revealed that multimorbidity affects more than 91% of people with cancer [53].

Moreover, Portugal was one of the countries with higher rate of multimorbidity[45].It can be proven in [46], suggesting a progressive increase in the aging of the Portuguese population at 2050, and therefore the phenomenon of multimorbidity. Furthermore, in 2015, Prazeres *et al.* [47]revealed that present in multimorbidity was present in 72.7% of Portuguese population, and Cardiometabolic and mental disorders were the most common chronic health problem .

Researches [41],[43] referenced the succeeding most prevalent diseases: a) Cardiovascular diseases (CVDs), especially Coronary heart disease (CHD), when arteries are obstructed and difficult the blood flow; b) Diabetes is a chronic disease in which blood sugar (glucose) levels are abnormal due to impaired insulin production, secretion, and action; c)cancer is characterised spread of abnormal cells; d) diseases of the airways, such chronic obstructive pulmonary disease (COPD), difficult air flows to and from the lungs; e) Arthritis which results in painful inflammation and stiffness in joints, limiting the ability to move.

Patterns

For a better understanding of the burden, determinants, prevention, and treatment of patients with multimorbidity, currently, the prevalence of specific combinations of multimorbidity is considered more than the prevalence of isolated chronic diseases.

To overcome the difficulty and complexity of this analysis, several approaches have been used to group multimorbidity into different combinations or patterns of comorbidities. Identifying how multimorbidity trends and patterns change over time will help clinicians predict the possible occurrence of multimorbidity risks among patients and recognize coexisting diseases. This is important as some diseases coexist more often than others. Therefore, it becomes necessary to demystify the patterns of multimorbidity, comprising highest concerning risk factors, and high hospitalizations and mortality[25], [37], [48].

Some of most common long term conditions (LTCs) include chronic obstructive pulmonary disease (COPD), heart failure, CVD, diabetes and cancer [19]. Conforming to some articles, the top three groups of multimorbidity patterns with relevant similarities are “Cardiovascular and metabolic diseases” , “Mental health problems” and “Musculoskeletal disorders”[25],[33].

Recent refinements in statistical approaches have resulted in improved methods to capture patterns of multimorbidity and detect subgroups, such as cluster analysis (cluster diseases) and latent factor analysis (patient grouping)[48]–[50]. Both types of methods are relevant as they allow discovering hidden relationships between conditions. In this sense, the careful evaluation of this information by clinicians and policymakers will be useful for strategies of prevention, diagnosis, treatment, prognosis, and adequate availability of resources [51].

1.3.4. Burden of diseases and healthcare services

Despite advances in public health, with improvements in clinical interventions and survival, many efforts are crucial to manage complex patients in primary care, with two or more comorbidities. For example, some guidelines were developed to assist this concern in multimorbidity. The National Institute for Health and Care Excellence (NICE) is a useful summary assessment and management of co-morbidities, which focuses on the need for patient-centred care [18], [49]. However, it loses long-term utility as it considers health problems as separate diseases. It should be noted that symptoms, frailty, limitation in daily activities, mood and dependence occur as a result of the various comorbidities.

Changes and/or progress of the chronic disease affects patients, especially in mental health, enabling them to perform daily tasks and manage their conditions. Moreover, disability or frailty declines the quality of life, so that, a rise of mortality [52],[53]. Larrañaga et *al.* [16] suggested the need to go beyond the simple

counting of chronic conditions and relate multimorbidity with an increased risk of functional decline. Globally, chronic diseases account for about 41 million deaths each year, equivalent to 71% of all death[54]. Also, Nunes *et al.* [55] revealed a positive association between multimorbidity and mortality.

Consequently, this represents a general growth in the burden of disease on health systems, concretely, in primary health care. In fact, long-term chronic conditions (LTCs) admit greater use of health services, including regular hospital visits, readmissions and emergency admissions[56]. All of them represent important outcomes to predict, in order to apply strategies regarding multimorbidity decision-making and disease management. Because of healthcare utilisation, there is a serious impact on costs. In accordance with [33] it is not veracious to estimate the cost for each patient as a sum of the costs for every single condition the patient has. From a health services and financing perspective, an evaluation of a patient's condition and outcomes of interactions among co-existing diseases makes more sense for cost estimation. Studies varies on healthcare system and methods used to derive it, however has been related an high rate of costs face to multimorbidity patients[44],[57],[58].

An additional concern is that patients with co-existing conditions are often referred to multiple medical specialities, leading to fragmented care, miscommunications, and other complications [38]. The information about patient profiles on hospital databases needs to be clear, for a more personalised and successful treatment.

As the prevalence and patterns of multimorbidity relate to social inequalities, policymakers also need to pay attention to certain aspects when developing policies to improve care for people with multimorbidity, namely inequity in access to health and social care, as well as in access and use of eHealth solutions[18].

An approach to developing appropriate health policies, mainly at primary care level, is the proper identification of risk factors, which corresponds to any characteristic or exposure of an individual that increases their likelihood of suffering from a disease or injury. As mentioned before, some chronic diseases can co-exist together by sharing them. Significant predictors of multimorbidity cover socioeconomic and demographic factors, together with lifestyle habits, such as smoking, alcohol intake, lower physical activity and poor diet [18],[33]. Nevertheless, ageing is the main one, associated with organ and system dysfunction, leading to a diminished response to different pathogens and to the development of a chronic inflammatory process.

In oncology, there is evidence that multimorbidity conditions impact every stage of cancer, such as stage at diagnosis, treatment, and outcomes of people with cancer. A first concern is that cancer symptoms may be mistakenly considered as symptoms of pre-existing health conditions, and could delay diagnosis [59]. Moreover, it is well-known that prognosis of gastric cancer is very poor, with a 5-year survival rate higher than 20% for advanced disease [60].

Overlook this circumstances, health institutions must develop strategies for providing integrated patient-centred care using conceptual and strategic vision, invest in the development of care professionals' knowledge and competences, enhance coordination and collaboration between institutions and organisations and supporting patients in their need, improving the responsiveness and resources to a personalised care[61].

1.4. Machine Learning and Healthcare

Medical databases contain relevant information for patient's care, such as demographics, diagnoses, medical procedures, medications, vital signs, immunizations, laboratory results, and radiology images. Heterogeneity of data presents challenges in storage of data, data integration, scalability, processing, visualisation and transmission, missing information, lack of standardisation, generalizability, and loss of quality. Also, huge information is generated at a high speed, becoming important to convert this data into useful information and knowledge. This particular field is named Data mining, which is the process of

discovery of hidden patterns and relationships in the data, resulting in knowledge and potentially actionable insights from the data[62],[63].

Data mining has been extended to expand classical statistical analysis methodologies, in particular by using machine learning (ML). This constitutes the main method which aims to identify patterns through algorithms. As already mentioned, the stratification of patients by multimorbidity patterns allows a more accurate understanding of the relationships between diseases and, therefore, a more targeted and tailored treatment. In this sense, ML will improve the quality of medical care, while optimising medical processes and management strategies [62],[64]. During the entire process, it is strictly necessary a collaboration and a mutual understanding between medical experts and data analysts, in order to prevent fragmented clinical care[18].

ML methods can be distinguished according to two main learning strategies: supervised learning (SV) and unsupervised learning (USV). The SV is implemented through a predefined set of classes, and its common tasks include a) classification, which comprises identification of classes in terms of their attributes and determining what class new items will belong to; b) regression analyses to visualise dependence between the values of the attributes and 3) prediction for forecasting the value of a specific attribute. Nonetheless, an eminent disadvantage is that the results of the test set only show that similar data can obtain similar results using the model, and do not guarantee model correctness when new datasets are considered [64],[65].

Unsupervised learning infers the underlying patterns in unlabelled data to find sub-clusters of the original data, to recognize outliers in the data, or to produce low-dimensional representations of the data. For that reason, this method is routinely applied in the context of multimorbidity[66]. Common unsupervised learning methods include principal component analysis (PCA), clustering and association analysis[64].

Essentially, PCA is used to denoise and to reduce dimensionality. It allows an overall "shape" visualisation of the data, by applying a linear transformation and traced a straight line that separates the directions of eigenvalues with the most substantial variation in the covariance matrix. From this separation results the “Principal components”. The succeeding decomposition lowers their eigenvalues, the variance decreases, and the high-dimensional set is simplified. Some limitations are related with the fact that PCA is only based on the mean vector and covariance matrix and only considers orthogonal transformations (rotations) of the original variables [67].

Cluster analysis is an unsupervised machine learning technique that handles the interaction of multiple variables to define subgroups of individuals with similar attributes. Another technique is Association Rules to identify which sets of diseases are frequently occurring together. This will be better explained in sections 1.3.1 and 1.3.2.

Furthermore, exploratory factor analysis is often used to identify multimorbidity patterns [22], [25], [68]. It allows for understanding the relationship between the items and discovering the underlying factors that the items may have in common. Factor loading is the correlation between the item and the factor[69].

An enhanced and intelligent technology that combines science mapping with performance analysis is network graphics. In the context of multimorbidity, this design is useful because it allows the inclusion of a large number of diseases, helping to identify highly connected chronic conditions and the magnitude of such connections. Each node represents a disease and a weighted edge between them reflects the strength of the relationship, by varying the thickness and colour. Blue and thicker edges correspond to strong statistical correlation positive relationships, which are the opposite of red edges. Sometimes, faced with a complex network, a meticulous analysis of each node can be costly and time-consuming. A possible approach for this issue is to apply a centrality measure to describe the importance of each node in the network, and determine the best one[25],[70],[71].

1.4.1. Clustering

Cluster analysis has been applied to a wide range of applications as an exploratory tool to enhance knowledge of no labelled data. This type of analysis corresponds to an iterated process of finding meaningful subgroups in a given dataset and associating data points with common properties [53]. In the context of multimorbidity, clustering of chronic conditions is crucial in understanding their most common combinations and determining outcomes on health and mortality. From here, patients within a certain cluster will be treated and handled differently to patients not belonging to that group. Hence, this procedure will help hospital managers to allocate limited resources, control patient costs effectively, and improve the quality of medical services[22],[37].

There are some basic steps to follow during a clustering task to achieve the optimal final solution. A preprocessing step is fundamental in order to:

1. Remove noise or outliers from data when data is sensitive to that
2. Data normalisation, which is important for distance-based clustering
3. Delete irrelevant attributes to accurate computational time.

Precisely, scaling data is fundamental to calculate distance differences between points in a dataset and allowing the grouping of the closest ones together or separate the furthest ones. The selection of distance measures will play an important role for further steps, including in obtaining correct clusters[72],[73].

The effectiveness of the method depends on the definition of (dis)similarity measures. In the literature, the most well-known distance used for numerical data is probably the Euclidean distance. It determines the sum of squared distance between two uncorrelated data points. Besides the easier computation, this is very sensitive to outliers. The Manhattan distance comes up from the Euclidean distance, and it considers horizontal and vertical components and determines the absolute difference among the pair of coordinates[72],[73].

Other measures consider some specific properties of the data points, instead of their specific location in a space. Jaccard index is a statistic used for comparing the similarity and diversity of sample sets, calculating the intersection of those items divided by their union. Cosine similarity focuses on the directional similarity. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. Lastly, the Gower distance is a metric that measures the dissimilarity of two items with mixed numeric and non-numeric data. That is useful to estimate a low-dimensional representation of high-dimensional data, while balancing the contribution of the different variables to the overall distance. It requires an NxN distance matrix to be calculated, for example a dissimilarity matrix. Particularly, the data are grouped according to the associated score that assumes the value 1 if the categories are the same and 0 if they are not. A summary table is presented below, along with commonly used distance measures, and their respective formulas [26]-[29],[74].

Depending on the study and dataset, there are different formulations of clustering: the partitional clustering, hierarchical clustering and Density-Based Methods. The first one produces a single-level clustering result, whereas hierarchical clustering outputs multilevel nested decompositions. The density-based methods identify the clusters as regions of high-density, which are separated by regions of low-density and, for that reason, the number of clusters doesn´t need to be prior fixed[73].

The aim of partitional clustering is segmenting a dataset into more homogeneous k clusters and can be divided in centroid or model-based methods. The centroid approach is characterised by a central point (e.g. the mean, the median, etc.) computed for the clusters, and for each iteration the points are assigned to a new position in relation to the centroid,[65],[72]-[74].

The K-Means algorithm is quite simple and fast to implement, but very sensitive to the initial centroid selection, because for different initial values, different clusters can be generated, and abnormal points disturb the mean value. In fact, the centroid may be shifted to a wrong position if the data has outliers. Also, despite

minimising intra-cluster dissimilarity, it does not ensure that the obtained solution is a global minimum [65],[72]-[74].

Table 2 – Common measures of similarities [26]-[29],[74].

Distance measure	Formula
Euclidean distance	$\sqrt{(x1-x2)^2+(y1-y2)^2}$
Manhattan distance	$\lvert x1-x2\rvert+\lvert y1-y2\rvert$
Jaccard Similarity Coefficient	$J(A,B)=\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$
Cosine Similarity	$q\ =\arccos \mathrm{A}\ \bullet\ \mathrm{B}\ \lVert A\rVert\lVert B\rVert$
Gower's Similarity Coefficient	$S_{Gower}(xi,xj)=\frac{\sum_{k=1}^{p} sijk}{p}$

Partitioning Around Medoids (PAM) constitutes a more robust version of the K-Means to noise and outliers. It can be explained by using the medoid as a reference point for the cluster creation, the most centrally located object in the cluster, instead of using the mean points as centroids. The **Figure 2** explains how means and medoids positions can vary in the presence of an outlier[75]. However, the most eminent disadvantage is related to the fact that PAM is computationally costly as it performs clustering on the overall data set. The CLARA method arises from PAM to handle large datasets, using only random samples of the input data, instead of the entire dataset[65],[72]-[74].

More recently, fuzzy c-means comes up as a soft clustering approach. It generalises the partition-based clustering method by allowing a data object to be part of more than one cluster. Fuzzy clustering is especially useful when there are no clear boundaries between clusters, however it is more complex and therefore more time consuming[67], [76].

Hierarchical clustering presents a particular representation, a multilevel hierarchy or dendrogram, allowing to see the best tree cut level for generating suitable groups. An advantage is that it doesn´t require any knowledge about the appropriate number of clusters beforehand. The input to a hierarchical clustering algorithm consists in the measurement of the similarity (or dissimilarity) between each pair of objects[67], [73].

Hierarchical methods are based on two strategies: (i) divisive, and (ii) agglomerative. The first one corresponds to a top-down clustering approach, starting with all observations which will be assigned to a single cluster. During the process, clusters will be paired by similarity until there is one cluster for each data or observation; an agglomerative approach considers each observation as a cluster itself and most similar clusters are successively merged into bigger clusters. Also, this are based on a linkage method, specifically, single and complete linkage measure the minimum and maximum distance between clusters, respectively; average linkage, which calculates the average of distances between all pairs, centroid method, combining clusters with minimum distance between the centroids of the two clusters; or finally Ward's method that minimises the total within cluster variance [65],[66],[72]

Since clustering lacks a priori information about structure in data, his analysis constitutes a challenge task. Because of that, researchers have been trying to interpret and evaluate results in a proper manner. Once the results have been obtained from the clustering, their validation is needed to obtain significance and confidence of clusters. External clustering validation and internal clustering validation are the two main categories of clustering validation, but while the first one uses information does not present in the data,

internal validation measures only rely on information in the data. There exist several criteria of adequacy, based on cluster properties, such as compactness, which measures how dissimilar are examples that lie within the same cluster, and separation, that determines the isolation of one cluster from the other [66], [67],[76].

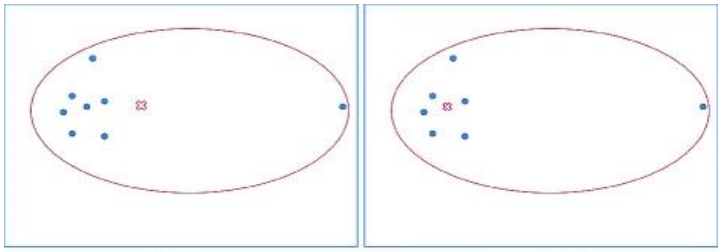

Figure 2 – (a)Mean vs (b)medoid in 2-D space. In both, the red point represents the centre. Image extracted from [74]

1.4.2. Association mining

Research on association rules started as early as the 1960s and is another USV method. There are two purposes, including to find frequent and infrequent item sets in a database, as well as finding associations, positive or negative, in the above sets. Particularly, they are stored in the form of transactions, with each transaction containing one or more items [65].

These cooccurrences are represented in the following form:

$$LHS \Rightarrow RHS\ [support,\ confidence]$$

, where the left-hand side (LHS) implies the right-hand side (RHS).

To evaluate the quality of each rule, there are two main measures, support, and confidence. The first includes all items in the antecedent and consequent parts of the rule. Confidence expresses how many transactions that include items from LHS, also include items from RHS. Both are specified, who defined minimum support and confidence. Rules having the support and confidence value greater than user specified values, are called valid association rules. Frequent item sets have traditionally been used to generate positive association rules; but that doesn't always happen [65]. Sometimes frequent items can be negatively correlated [77]. In this sense, the lift measure is considered to characterize the direction of the relationship between the antecedent, *LHS,* and the consequent, *RHS*. A lift value greater than 1 indicates a positive relationship between the item sets; and a negative one indicates a negative relationship. When lift is equal to 1, the item sets are independent and there is no relationship between the item sets [77],[78]. Ideally, the goal would always be to discover rules that have 50% support and 90% confidence, but unfortunately, the corresponding number of rules would be very small or even null.

1.3. State of art

For some years, the lack of a harmonised definition of multimorbidity, including number and types of conditions and the measure method, has created difficulties to compare this condition across studies and explore its impact[30].

In 1998, a team [33] decided to set 30 comorbidities to test their heterogeneity between different groups of patients. As a result of the algorithm applied, the comorbidities were associated with substantial increases in length of stay, hospital charges, and mortality, for all groups. However, the lack of distinction between independent comorbidities and conditions that are directly related to the principal diagnosis contributes to overestimate the contribution of comorbidities on health outcomes. In 2021, a systematic review was performed by Ho et *al*. [24] to assess which conditions 566 international studies included in the measure of multimorbidity. The studies were peer reviewed, and it was noticed that the number of conditions embraced in the measurements ranged from two to 285. Considering just one comorbidity, successive conditions were highlighted, in decreasing order of percentage: cardiovascular, metabolic and endocrine, respiratory, musculoskeletal, and mental health. Filtering out the chronic diseases included in more than half of the studies on the multimorbidity measure used, the group highlighted Diabetes, Stroke, Cancer, Chronic Obstructive Pulmonary Disease, Arterial Hypertension, Coronary Artery Disease, Chronic Kidney Disease, and Heart Failure. As a consequence of this work, a few months later, the same author decided to investigate what factors contribute to the heterogeneity of multimorbidity prevalence. [39] Across the 193 studies, they found a large variation in the number of conditions included in multimorbidity measures (ranging between 2 to 285 conditions), and which conditions were included. Hence, the mean age and number of conditions were the main predictors for that measure (highest prevalence for ≥74 years and ≥44 conditions). Despite this, only eight conditions were included in more than half of the studies, which were similar to the previous study. New insights were revealed, contemplating hypertension, diabetes, and cardiovascular disease (CVD) as often coexisting chronic diseases with cancer.

As previously noted, multimorbidity research has generally referred to a count of chronic conditions. By virtue of that, Dekhtyar *et al.* [40] examined the association between lifelong experiences and the rate of accumulation of chronic diseases in old age. Through mixed linear models, life experiences and the number of diseases present were assessed over time. Higher education, lifelong active occupations and richer social networks decrease multimorbidity after age 60. In 2011, Diederichs *et al.* [32] introduced a new approach to weighted multimorbidity indices. A revision between 1960 and 2009 containing in-patient medical records, investigated different diseases as an indicator of multimorbidity. Based on self-reported health status they found 11 conditions as the most common causes (including cancer, depression, myocardial infarction, and hypertension) of inpatient and outpatient care, as well as death in people over 64 years in Germany. In a hospital context, it is crucial to measure the presence of multimorbidity in the use of health services and hospital mortality. There are a large variety of studies exploring mortality outcome as a consequence of multimorbidity. In 2012, a guide made by Huntley et *al*[30] revealed that for assessing care utilization the best indexes were ACG System, the Charlson' index, or disease counts. Particularly, the Charlson' index was also presented as a mortality predictor. Also, a Cezard et *al*. [32] review, presents Linear mixed models as the major approach to assess the association between predictors (e.g. life experiences, age and ethnic variations, number of comorbidities, etc) and the speed of multimorbidity accumulation and mortality.

Over time, multimorbidity analysis evolved into more sophisticated mathematical techniques, shifting the focus from the disease to the patient. In this sense, several studies began to consider each individual as an analysis variable in the assessment of multimorbidity patterns.

Investigating clusters of combinations of conditions within a population rise up as a development to integrated care models. One of the most known studies was published in 2012, by Torres *et al.* [13], in which factor analysis was used to find patterns of multimorbidity. From records of 19 primary care centres from 2008, there were identified five patterns of multimorbidity according to common underlying causal factor: 1) cardio-metabolic, 2) psychiatric-substance abuse, 3) mechanical-obesity-thyroidal, 4) psychogeriatric and 5) depressive, with a prevalence ranging from 13% in the 15 to 44 years old to 67% in those with 65 years of age

or older. Agterberg *et al.* [34] created clusters based on demographics, economic and health-related conditions of individuals. A comparison between k-means and PAM was made to identify clusters with highest health expenditures, an important predictor in health multimorbidity outcomes. Also, a study using electronic medical records from the UK between 2005–2016, used statistical methods and network analysis to identify comorbid pairs and triads of diseases and to identify clusters of chronic conditions across different demographic groups [79]. By virtue of network analysis, the most meaningful clusters across all demographics were respiratory, cardiovascular and a mixed cardiovascular-renal-metabolic cluster. Moreover, Bisquera *et al*. [49], selected UK patients older than 18 years , between 2005 and 2020 to distinguish groups of multimorbidity from 32 long-term conditions (LTCs). Through multiple correspondence and cluster analysis, five consistent LTC clusters were detected: 1) anxiety and depression; 2) heart failure, atrial fibrillation, chronic kidney disease (CKD), chronic heart disease (CHD), stroke/transient; ischaemic attack (TIA), peripheral arterial disease (PAD), dementia and osteoporosis ;3) osteoarthritis, cancer, chronic pain, hypertension and diabetes; 4) chronic liver disease and viral hepatitis; 5) substance dependency, alcohol dependency and HIV. A complete study occurred in 2015, by Held *et al.* [80], to evaluate multimorbidity in people aged 70 years old or more, using network analysis and Association Rules methodology . The first one showed five clusters, including vascular, metabolic, neurodegenerative, mental health and other, and a musculoskeletal and other. Association rules originated 18 dyads and 1 triad of morbidities which reflected the overall prevalence of these morbidities in the study population, with arthritis with hearing impairment being the most common dyad, and the most common triad was arthritis with hearing impairment and obesity. Three years later, Llorach *et al.* [33], applied hierarchical cluster analysis and exploratory factor analysis to records of 408 994 patients with multimorbidity aged 45–64 years, in Catalonia primary health care. To sum up, the team denoted that Exploratory factor analysis (EFA) could be more useful to analyse comorbidities and describe the correlation between diseases that have a pathophysiological relationship, while Hierarchical Clustering analysis (HCA) will be useful in disease associations with random coexistence of diseases or without causal explanation. Another article analysed multimorbidity patterns, describing a large variability in their prevalence among European countries [81]. Related to the number of chronic diseases, most of the study population reported being diagnosed with two chronic diseases. In younger women, mental and osteoarticular disorders are more prevalent when compared to younger men, and stroke and diabetes are more prevalent conditions in men and increase with age. The patterns vary on studies by statistical techniques used, such as cluster or factorial analysis, or by analysing the frequency of combinations of the major diseases. More recently, Robertson *et al.* [82] characterised multimorbidity clusters among adults admitted to a hospital in Grampian, Scotland, in 2014, and who had ≥ 2 of 30 chronic conditions diagnosed in the 5 previous years. Operating Gower distance and Partitioning around Medoids, ten clusters of similar conditions were identified: hypertension, asthma, alcohol misuse, chronic kidney disease and diabetes, chronic kidney disease, chronic pain, cancer, chronic heart failure, diabetes, and hypothyroidism. Also, in 2021, Lee *et al.* [83] admitted the importance of identifying the prevalence and patterns of multimorbidity among Korean adults . Association rule analysis was performed on pairs of diseases with high prevalence and network analysis was conducted to identify the association between fifteen frequent diseases among men and women. In men, the most relevant associations were Diabetes with Hypertension and Dyslipidemias with Hypertension. Women also shared a high percentage of Dyslipidemias with Hypertension with men but differed in higher Polyarthrosis-Hypertension relationship. Contrary to the previous analysis, the results of the network analysis were divided into four groups by gender and age. Hypertension and dyslipidemia had a high degree of centrality in all groups. Specifically, in men aged under 65, Gastritis and allergic disease had a high degree of centrality, while above 65, the most common diseases were Prostatic Hyperplasia, diabetes mellitus and cataract. In younger women, gastritis, polyarthritis, and disc disorder had a high degree of centrality, whereas older groups presented mostly polyarthritis, osteoporosis, and cataract.

For last, a study examines associations between upper gastrointestinal cancers and 50 long-term conditions [84] by applying Least Absolute Shrinkage and Selection Operator (LASSO). This serves to model the relationship between long-term conditions as predictors for the dependent variable, the upper gastrointestinal cancer. As a result, participants with bronchiectasis, diabetes, Parkinson's disease and psoriasis/eczema demonstrated a greater risk of oesophageal cancer, while participants with chronic fatigue syndrome, glaucoma and multiple sclerosis observed a greater risk of stomach cancer.

1.4. Main aim and objectives

The main goal of this thesis is to apply unsupervised machine learning techniques, namely cluster analysis and association rules, in alignment with the literature, in order to identify and describe multimorbidity groups in Portuguese nationwide stomach cancer hospitalizations, as well as investigate the effects of multimorbidity on health outcomes for this group of patients.

In this sense, the following specific research questions will be resolved:

1. Is clustering analysis suitable to detect clinically relevant multimorbidity clusters among stomach cancer hospitalizations?
2. What are the most common pairs, triads, or even more groups of co-existing comorbidities among stomach cancer hospitalizations in Portugal?
3. What is the effect of multimorbidity on in-hospital mortality and length of stay among stomach cancer patients and how it interacts with other patient characteristics, such as demographics, cancer type and presence of metastasis, in the determination of both outcomes?

2. Materials and Methods

2.1. Study population, sample selection and definition of variables

For the elaboration of this dissertation, administrative data provided by the central administration of the health system (ACSS) was used, containing data on all stomach cancer hospitalizations that occurred in public hospitals in mainland Portugal, between 2011 and 2015. All episodes presenting International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes of stomach cancer (151.X) as the main diagnosis were filtered, as shown in the **Table 3**.

The selected sample has the following variables:

- **Region**: nominal qualitative variable with six categories that identifies the region of Mainland Portugal (Norte, Centro, Lisboa, Alentejo and Algarve) where the patient lives.
- **Sex**: nominal qualitative variable with two categories, corresponding to the sex of the patient:
    1. Male
    2. Female
- **Age**: continuous quantitative variable of the user's age, am years, later converted in a categorical variable comprising three main groups "[0-49] years", "[50-74] years" and "[>75] years"
- **Length of stay** (LOS): discrete quantitative variable, corresponding to the total number of days spent by the patient at the health institution, considering that hospitalisation presents a minimum of 24 hours of stay in the hospital.
- **Main diagnosis**: nominal qualitative variable of the ICD-9-CM code, which identifies the main diagnosis. In this particular case, it was filtered to 151.0, 151.1, 151.2, 151.3, 151.4,151.5, 151.6 (See **Table 3**)
- **Comorbidities**: binary variable that identifies the presence or absence of 17 comorbidities identified according to the Charlson' method. Charlson' comorbidities were identified according to the definitions based on the work of Quan *et al.* [27]. Specifically, the following variables regarding each Charlson' comorbidity were included in this work: Myocardial infarction (MI), Congestive heart failure (CHF), Peripheral vascular disease (PVD), Cerebrovascular disease (CVD), Dementia (DEM), Chronic obstructive pulmonary disease (COPD), Rheumatologic disease (Rheum), Cerebrovascular disease(CEVD), Peptic ulcer disease(PUD), Mild liver disease (MILDLD), Moderate or severe liver

disease(MSLD), Diabetes mellitus without chronic complications (DIAB_UC), Diabetes mellitus with chronic complications (DIAB_C), Moderate-to-severe renal disease(RD), Paralysis (PARA), Cancer without metastases (CANCER) and HIV(HIV). Although the CCI typically includes metastatic cancer (METS) diagnoses, this condition was excluded because it may be an extension of the cancer of interest.

- **Multimorbidity**: presence of two or more Charlson' comorbidities

Table 3 - ICD-9-CM Coding Algorithms for cancer of stomach

ICD-9-CM code	Categories of Cancer of stomach
151.0	Malignant neoplasm of cardia
151.1	Malignant neoplasm of pylorus
151.2	Malignant neoplasm of pyloric antrum
151.3	Malignant neoplasm of fundus of stomach
151.4	Malignant neoplasm of body of stomach
151.5	Malignant neoplasm of lesser curvature of stomach, unspecified
151.6	Malignant neoplasm of greater curvature of stomach, unspecified

2.2. Software

R is a free software environment, which presents a variety of useful packages for statistical computing, graphics visualization, and machine learning. Its application in the health area has been increasing due to the easy use of technique to hospital and clinical data analysis. During the present work, R version 4.2.0 for Windows was requested.

2.3. Data treatment

In this step were performed the necessary initial transformations such: a) convert strings and numerical values to factors; b) allocate random "ID" for each patient c) subset different types of stomach cancer and create a new database, for each one, for an individual analysis; d) create two different groups based on stomach anatomy: Group 1 comprises codes 151.0 + 151.3 +151.4 (cardia/fundus/body), and Group 2 with codes 1511 + 1512(pylorus/pyloric antrum).

2.4. Frequency and relative frequency of socio-demographic variables and Charlson' comorbidities

In order to create a frequency and relative frequency table, a combination of various functions, from "dplyr" package [85], **group_by (), summarise (), n (), mutate (), and sum (),** were applied in original database("dados"). Hence, unique values of three variables, age, sex, and Charlson' comorbidities were counted to summarise frequency and obtain relative proportions, as following:

```
dados %>% group_by(age)%>% summarise(n=n()) %>% mutate(freq = (n/ sum(n)*100))

dados %>% group_by(sex)%>% summarise(n=n()) %>% mutate(freq = (n/ sum(n)*100))

dados %>% group_by(comorbidity)%>% summarise(n=n()) %>% mutate(freq = (n/ sum(n)*100))
```

2.5. Prevalence of multimorbidity in study population

The first step in this section was to count the number of comorbidities for each event/patient hospitalization using **mutate ()** function, from "dplyr" package, as *mutate (count = rowSums(. == 1)),* and further filtered using *count* ≥ 2. Age and sex variables were selected by select () function, from "dplyr" package, as following *dados %>%* ***select*** *(sex, age).* Using **cbind ()** approach, a new dataframe was created with three columns (age,sex,n_diseases), where n_disease a column obtained in 2.4. with the number of chronic diseases counted and filtered for two or more Charlson' comorbidities, for each event/hospitalization. To condense only values respected to age and sex into summary statistics, two lines of code were established operating **tbl_summary ()** from "gt_summary" package[86],as following:

multimorbidity %>% tbl_summary(by = count)

2.6. Prevalence of socio-demographics conditions for each malignant neoplasm of stomach in multimorbid patients

The next step included creating a database for each type of malignant stomach neoplasm of stomach (*malignant neoplasm of body of stomach, malignant neoplasm of pyloric antrum, malignant neoplasm of lesser curvature of stomach, unspecified, malignant neoplasm of cardia, malignant neoplasm of pylorus, malignant neoplasm of fundus of stomach, malignant neoplasm of greater curvature of stomach, unspecified*) identified in the multimorbid Portuguese population between 2011 and 2015. The proportion of patients of "female sex" or "male sex" and "[0-49] years old", "[50-74] years old" and "[>75] years old" was calculated from **group_by (), summarise (), n (), mutate (), and sum ()** functions again, as shown below:

*Prev_age <- type1%>% group_by(age)%>% summarise(n=n()) %>% mutate(freq = (n/ sum(n)*100))*

*Prev_sex <- type1%>% group_by(sex))%>% summarise(n=n()) %>% mutate(freq = (n/ sum(n)*100))*

2.7. Dissimilarity matrix

After an exclusive section of patients with two or more comorbidities, the dissimilarity matrix was calculated to find the difference between each chronic Charlson' comorbidities distance.

To maintain meaningful results, in the present work categorical data has been maintained, and Gower's similarity coefficient was applied to the dissimilarity matrix. For this purpose was applied **daisy ()** function, from "cluster" package [87], as follows: *daisy(x, metric = "gower"*), where x is dataframe, and dissimilarities are computed between his rows. Those will be inputs to cluster analysis and multidimensional scaling.

2.8. Clustering

In this dissertation, the first diseases 'grouping approach based on their similarities was clustering, which is divided in two principal categories: partitional clustering and hierarchical clustering. On the one hand, a partition normally requires an initial number of clusters k as an input parameter, that will be posterior submitted to iterative optimization. On the other hand, hierarchical clustering is a set of nested clusters that are arranged as a tree.

2.8.1. Clustering implementation

In this section there are showed the general steps of each clustering algorithms.

K means and PAM

1. Choose initial k number of cluster centers(centroids) randomly or based on some prior knowledge.
2. Calculate their distance from all the points in the scatter plot;
3. Classify each point into the cluster whose center it is closest to;
4. Select a new point in each cluster that minimizes the sum of distances of all points in that cluster from itself;
5. Repeat Step 2 until the centers stop changing.

The PAM algorithm is identical to the k-means clustering algorithm, except for Step 1 and Step 4. The center of PAM cluster is a medoids, always a data point in the dataset. In Step 4, the difference is that PAM minimizes the sum of dissimilarities instead of a sum of squared Euclidean distances. This is a more robust approach.

Hierarchical

Hierarchical clustering is a method to group data into a tree of clusters. **H.clust ()** computation was applied, using the dissimilarity matrix. Then, it repeatedly executes the subsequent steps:

1. Each object is assigned to its own cluster;
2. Several iterations, at each stage joining the two most similar clusters, continuing until there is just a single cluster.

Particularly, in the present work was used Ward's minimum variance method to find compact and spherical clusters.

2.8.2. Determining the optimal number of clusters

The partitioning methods, k-means and PAM, require an initial parameter of the number of clusters. There is no best procedure to obtain it, because it depends on the method used for measuring similarities and the parameters used for partitioning. The direct methods comprise within cluster sums of squares or the average silhouette, instead of gap statistic, a statistical testing method [76].

- **Elbow method** is an approach not merely to find the optimal number of clusters, but also to interpret and validate the consistency within clusters of data. Basically, the silhouette coefficients of each point are computed to measure how much a point is similar to its own cluster compared to other clusters[88];

- **Silhouette coefficient** is an approach not merely to find the optimal number of clusters, but also to interpret and validate the consistency within clusters of data. Basically, the silhouette coefficients of each point are computed to measure how much a point is similar to its own cluster compared to other clusters.

The silhouette coefficient is calculated as follows:

$$s(p) = \frac{(b(p) - a(p))}{max\ \{a(p), b(p)\}}$$

, where a(p) is the mean distance between point p and all other points within the same cluster (intra-cluster distance), whereas b(p) is the smallest mean distance of p to all points in any other cluster, where p is not a member (nearest-cluster distance).

The coefficient can take values in the interval [-1, 1]. If it is a) 0 the sample is very close to the neighbouring clusters; b) if it is 1, the sample is far away from the neighbouring clusters; and 3) if it is -1, the sample is assigned to the wrong clusters.

- **Gap statistic** can handle data that have a distribution with no obvious clustering (eg, globular, Gaussian, and slightly disjoint data distributions). comparing evidence against null hypothesis. The optimal value of the number of clusters occurs when the statistic is far from the uniform random distribution of points, so the ideal is the maximum gap statistic.

A first application, to determine the optimal number of clusters, was to apply fviz_nbclust (x, FUNcluster, method = c("silhouette", "wss", "gap_stat")), a function inside the package "factoextra" [85], where x is the dissimilarity matrix, and FUNcluster the clustering algorithm to be tested, namely k-means, PAM, and hcut (for hierarchical clustering). Both Elbow and Silhouette were performed, and the optimal number of clusters was inspected based on each corresponding representation: the number of clusters k versus WSS score and the number of clusters k versus Average Silhouette Width, respectively. Afterward, more accurate procedures were implemented, because give statistical insights instead of visualization requirements for finding the optimal number of clusters. Both functions calculate validation measures for a given number of clusters and clustering algorithms, so they will be better explained in 2.6.3.

2.8.3. Validation measures

In the present work, the cluster validation was used to obtain two results:

1. Determine the optimal number of clusters for every single type of cancer of the stomach and each algorithm.
2. Comparison between clustering algorithms and selection of the more reliable algorithm for the dataset.

An initial approach for 1 and 2 will be applied, using the **clust.stats ()** function, requires “stats” package [89], and “fpc” package [90], to evaluate the internal statistical properties, such as compactness and separability between points. The validation criteria includes the average silhouette width(avg.silh.width) and Dunn Index(dunn). As cited before, silhouette sets how well a point is fixed to its cluster compared to others, while Dunn index is a measure of the ratio between the smallest distance along observations not in the same cluster to the largest intra-cluster distance. It has a value between 0 and infinity and should be maximised. In the present study, the results were based only on the average silhouette width [49],[68].

In 1, decisions about the number of clusters were based on the maximum value of avg.silh.width between the two methods. Using the results of 1, the metrics will be compared between k-means, PAM and hierarchical clustering, in order to choose the most appropriate algorithm for each diagnosis of stomach cancer.

Considering the lack of consistency, given by the variation in results on different runs of algorithms, **clValid ()** function was applied to corroborate the results of **clust.stats ()** function. The package “clValid” [91], constitutes an enhanced approach compared to individual inspection of Silhouette and Dunn index for each algorithm. In a small piece of code, this function allows the user to simultaneously evaluate several clustering algorithms, number of clusters and determine the most appropriate method. It can compute in a single line function call, and compare standard algorithms, as following:

clValid(gower_mat, nClust = 2:5, clMethods = "kmeans", "pam" , "hierarchical", validation = "internal".

The measure selected was “internal”, which uses intrinsic information in the data to assess the quality of the clustering. Particularly, the number of clusters tested was set to vary only from 2 to 10, because too many clusters became less interesting and informative under the clinical and managerial perspective.

2.9. Association Rules

The purpose of this step is the application of Association Rules is to identify meaning combinations of diseases based on co-existence and not causality. The “arules” package[92], has been used in research to discover interesting relationships between variables in large databases. Between the two algorithms for mining association rules, apriori (Agrawal and Srikant, 1994) and eclat (Zaki, 2000) [93], the first one was selected in the present work. **Apriori ()** function only collects items that satisfy the minimum support requirement, in order to reduce the size of candidate sets and the cost function.

The following steps was used in **Apriori()** algorithm:

1. Categorical values are transformed into binary data
    - This step originates a transaction that only includes the subscripts (items) of binary dimensions equal to 1, this means the presence of chronic disease
2. Convert dataframe with categorical attributes into transactions
    - Categorical attributes are transformed into individual items, mapping each categorical value to one item. Each transaction is a set of items, and each item corresponds to the presence or absence of one categorical value.
3. Use object of class transaction as an input of **Apriori()** function
4. Select a threshold to support and confidence and find all the rules that exceed these
    - The minimum support was chosen to 2%, and the minimum confidence was set at 50%.

Apart from the specified minimum support and minimum confidence, all parameters have the default values.

2.10. Patterns visualization by Networks analysis

Next, a multimorbidity networks will be constructed to study the natural clustering of diseases in the dataset. The comorbidities were represented as nodes, and edges connect each pair of them, proportional to the strength of the association between the comorbid pair.

To estimate the network structure, **estimateNetwork ()** function was applied, from “bootnet” package [94], as follows:

estimateNetwork(data, default = "none", "EBICglasso", "pcor","cor").

Particularly, the data were all database of different types of stomach cancer and groups, converted to numeric type. By default, made use of “*pcor*” estimated a partial correlation network, grant a unique correlation between two variables that, when connected, their correlation cannot be explained by any other variable[95]. Distinctly, in the present work, correlation between comorbidities does not mean that one variable causes the other to occur (causality).

2.11. Estimation of impact of socio-demographic determinants and other conditions on outcomes

For a better clinical and health services management in multimorbidity, it is essential to evaluate the impact of additional comorbidities on relevant health outcomes, such as length of stay (LOS), which is a proxy of resource consumption, and in-hospital mortality.

Data sets in health care often fall outside the scope of basic statistics methods, which rely on normally distributions. Instead, health data are often binary (e.g., outcome is present or not present), proportions (e.g., mortality rates) or counts (number of days spent at the hospital). In basic statistical methods, such as linear models, the exact effects of each predictor variable are quantified, where variables between individuals do not change or change at a constant rate over time. However, in this type of models, the use of "fixed effects" may erroneously correct variables that may affect the result of the analysis, as health care problems often involve random effects, whose purpose is instead to quantify the variation by also encompassing variation among individuals (e.g., when multiple responses can be measured per individuals or groups, regions or time periods). Therefore, an adaptation emerges to linear mixed models, which also incorporates random effects and residual noise[96]. Attending to the nature of the present database, with non-normal data, Generalized linear mixed models (GLMMs) were implemented in order to quantify the effects of multimorbidity, controlled for other demographic and clinical variables, on in-hospital mortality and length of stay (LOS). Those statistical models combined mixed models with generalized linear models (GLMs), including normal, binomial, Poisson, and multinomial distributions as special cases[96].

For both outcomes, the effects of multimorbidity were assessed considering overall stomach cancer hospitalizations, different types of cancer according to the areas of stomach affected by the malignant neoplasm (cardia, pylorus, pyloric antrum, fundus of stomach and body of stomach). Furthermore, under a clinical and treatment management point of view, it is relevant to compare the effects of multimorbidity on patients with proximal cancers (cardia, fundus of cancer and body of stomach) versus those with distal cancers (pylorus and pyloric antrum). Particularly, the models implemented to quantify these effects on mortality and LOS included a random intercept for each hospital present the dataset, as we assumed that variation between hospitals is relevant in the context of our problem. Binomial and Poisson distributions were assumed to model mortality and LOS, respectively. Fixed effects or predictor variables considered for both outcomes were 1) multimorbidity level, which is a categorical variable that divides the number of comorbidities registered per episode into four groups, defined according to the distribution of comorbidities in the dataset: "0 comorbidities", "1 comorbidity", "2 comorbidities", "3 comorbidities", "4 or more comorbidities"; 2) A binary variable indicating the presence or absence of the metastatic tumour; 3) age group, represented by a categorical variable indicating "18-49 years old", "50-74 years old" and "75+ years old"; 4) a categorical variable indicating sex; 5) a categorical variable indicating the geographic region of residence of the patient, with five possible values ("Alentejo", "Algarve", "Centro", "Lisboa e Vale do Tejo" and "Norte"), as defined Nomenclature of territorial units for statistics (NUTS) 2 for Portugal [97].

The general form of the model can be described by the following equation:

$$\mathbf{y} = \boldsymbol{\beta_0} + \sum \boldsymbol{\beta_i x_i} + \gamma + \boldsymbol{\varepsilon}\,,$$

where y is the dependent variable, $\boldsymbol{\beta_0}$ is a global intercept; $\boldsymbol{x_i}$ corresponds to all variable with fixed effects; $\boldsymbol{\beta_i}$ is a change cause for each fixed variables; γ is variance resulting by random effect; ε is a portion of the residuals from normal distribution

Fixed effects are estimated, including the value of the t statistics (Student test), shown without the p-value, and the 95% confidence interval (CI), as:

$$CI = Estimate \pm 1.96 * Std.ErrorCI = Estimate \pm 1.96 * Std.Error$$

More formally, we estimate a two-level random intercept model, where we included several predictor variables measured at patient level, in order to identify the variation in mortality and LOS due to patient demographic (age, sex, region) and clinical characteristics (presence of metastatic tumor and

multimorbidity), systematic variation associated with the hospital where the hospitalization occurred, and random chance variation.

Regarding in-hospital mortality, which is a dichotomous outcome, a logistic random effects regression model was estimated (binomial distribution). The model can be written as follows:

$$logit(p_{eh}) = \beta_0 + \beta_1 x_e + v_h$$

, where p_{eh} is the log-odds of death for patient *e* , in hospital *h*. The intercept β_0 represent the log-odds of death when all predictors (which are categorical variables) assume the reference value across all episodes, whereas β_1 represents the effect in terms of changes in the log-odds of the demographics and clinical characteristics in relation to the reference value, at episode level. The exponential of these β_1 coefficients can be interpreted as odds-ratios. In the random part of the model, v_h is the effect of hospital *h*, and the variance of v reflects the degree of heterogeneity in mortality across hospitals[98].

Regarding LOS, a Poisson regression with random effects was estimated, which is a nonlinear model regression that is often used for modelling count data (in the case of the present study, LOS represented as number of days). The model can de written as follows:

$$log(u_{eh}) = \beta_0 + \beta_1 x_e + v_h$$

The intercept β_0 represent the baseline number of days when all predictors assume the reference values. The fixed part of the model captures the variation in LOS that is associated with the observable demographics and clinical characteristics of the patients. The coefficient β_1 indicates the difference in the log expected number of days (LOS) in relation to the reference category. The exponential of these coefficients can be interpreted as incidence rate ratios. The random part includes the observed variance between levels, including that attributable to hospitals, and that attributable to chance (i.e., not explained by the observed demographic and clinical characteristics)[98].

We further analyzed the residual variances that emanated information on the extent of variability in LOS at different hierarchical levels (episode and hospital). The proportion of explainable LOS variation attributable to each level was analyzed through the Intraclass correlation coefficient (ICC)[99]. The ICC was obtained between the variance at the hospital level and the total variance, and the results can be interpreted as the correlation (similarity) among observations within the same class (hospital). If the value of ICC is large, it can indicate that a considerable residual variance regarding LOS exists in that level. The ICC coefficient estimation method considering Poisson distribution can be found in [99].

For each assessed outcome, three models were estimated:

- **Model 1**: estimated for a subsample of episodes with a principal diagnosis of a proximal cancer (malignant neoplasm of cardia, fundus of cancer and body of stomach). Predictor variables included age group, sex, region, presence of metastasis and multimorbidity level.

- **Model 2**: estimated for a subsample of episodes with a principal diagnosis of a distal cancer (malignant neoplasm of pylorus and pyloric antrum). Predictor variables included age group, sex, region, presence of metastasis and multimorbidity level.

- **Model 3**: estimated for a subsample of episodes with a principal diagnosis of both, distal and proximal cancer. It included a categorical variable indicating whether the episode comprises a distal or proximal cancer, in addition to age group, sex, region, presence of metastasis and multimorbidity level.

3. Results

3.1. Frequency and proportion of relevant variables in ACSS database

This section presents a table with frequency and relative frequency of the categorical variables sex, age group, and Charlson' comorbidities. In particular, the first column describes the number of observations for each variable, while the second column shows the number of times a variable occurs compared to the total number of events.

Table 4 - Frequency and proportion of relevant variables categorical variables sex, age group, and Charlson' comorbidities for original database.

		Frequency(n)	Relative Frequency (%)
Multimorbidity	0 or 1 comorbidities	18425	89.0
	2 or more comorbidities	2230	11.0
Age	0-49 years	1682	8.14
	50-74 years	10954	53.0
	75+ years	8019	38.8
Sex	M	12514	60.6
	F	8140	39.4
Charlson comorbidities	MI	467	2.26
	CHF	1095	5.30
	PVD	394	1.91
	CEVD	661	3.20
	DEM	187	0.91
	COPD	1489	7.21
	Rheum	82	0.40
	PUD	305	1.48
	MILDLD	770	3.73
	DIAB_UC	3412	16.5
	DIAB_C	239	1.16
	PARA	69	0.33
	RD	760	3.68
	CANCER	686	3.32
	MSLD	152	0.74
	HIV	23	0.11

Myocardial infarction (MI), Congestive heart failure (CHF), Peripheral vascular disease (PVD), Cerebrovascular disease (CVD), Dementia (DEM), Chronic obstructive pulmonary disease (COPD), Rheumatologic disease (Rheum), Cerebrovascular disease(CEVD), Peptic ulcer disease(PUD), Mild liver disease (MILDLD), Moderate or severe liver disease(MSLD), Diabetes mellitus without chronic complications (DIAB_UC), Diabetes mellitus with chronic complications (DIAB_C), Moderate-to-severe renal disease(RD), Paralysis (PARA), Cancer without metastases (CANCER) and HIV(HIV)

Clearly, there is a higher proportion of male (60.6%) stomach cancer patients aged 50 to 74 years (53.0%). As regards Charlson' comorbidities, a large number of stomach cancer patients are visible who have Uncomplicated Diabetes (DIAB_UC) (16.5%), Chronic Obstructive Pulmonary Disease (COPD) (7.21%),

Congestive heart failure (CHF) (5.30), Mild liver disease (MILDLD) (3.73%), Moderate-to-severe renal disease (RD) (3.68%), and Cerebrovascular disease (CEVD) (3.20%).

3.2. Overall prevalence of multimorbidity for socio-demographic characteristics

The analysis was proceeded according to the number of comorbidities (0,1,2,3,4,5 or more), in which the multimorbid condition comprises two or more comorbidities.

Table 5 - Counted number and proportion of Charlson' comorbidities (0CHD: 0 comorbidities; 1CHD: 1 comorbidity; 2CHD: 2 comorbidities; 3CHD: 3 comorbidities; 4CHD: 4 comorbidities; ≥ 5ChD: 5 or more comorbidities) of stomach cancer patient for sex variable ("Male" and "Female")

	0 CHD (N = 12952)	**1 CHD** (N = 5473)	**2 CHD** (N = 1614)	**3 CHD** (N = 439)	**4 CHD** (N = 126)	**≥ 5CHD** (N = 51)
			SEX			
MALE	7642 (59%)	3422 (63%)	1044 (65%)	279 (64%)	92 (73%)	35 (69%)
FEMALE	5309 (41%)	2051 (37%)	570 (35%)	160 (36%)	34 (27%)	16 (31%)

Table 6 - Counted number and proportion of Charlson' comorbidities (0CHD: 0 comorbidities; 1CHD: 1 comorbidity; 2CHD: 2 comorbidities; 3CHD: 3 comorbidities; 4CHD: 4 comorbidities; ≥ 5ChD: 5 or more comorbidities) of stomach cancer patient for age variable ("[0-49] years", "[50-74] years" and "[>75] years

	0 CHD (N = 12952)	**1 CHD** (N = 5473)	**2 CHD** (N = 1614)	**3 CHD** (N = 439)	**4 CHD** (N = 126)	**≥ 5CHD** (N = 51)
			AGE			
0-49	1446 (11%)	203 (3.7%)	29 (1.8%)	4 (0.9%)	0 (0%)	0 (0%)
50-74	7253 (56%)	2787 (51%)	704 (44%)	155(35%)	38 (30%)	17 (33%)
75+	4253 (33%)	2483 (45%)	881 (55%)	280(64%)	88 (70%)	34 (67%)

From the two tables above, it appears that the majority of Portuguese patients with stomach cancer, between 2011 and 2015, had none (N = 12952) or one comorbidity (N = 5473). In the presence of multimorbidity, most patients with stomach cancer have two chronic diseases (N = 1614), with the proportion being higher for men (65%), compared to women (35%), and also higher for age. over 75 years (55%), compared to ages between 50 and 74 years (44%) and between 0 and 49 years (1.8%).

3.3. Influence of socio-demographic characteristics in stomach cancer types

From the point of view of hospital management, it makes sense to study the prevalence of sociodemographic characteristics for each type of malignant neoplasms of stomach cancer, assuming the presence of multimorbidity condition (≥2 comorbidities). Distinctly, the **Figure 3** is a barplot showing the sex (Male or Female) prevalence among different malignant neoplasms of stomach, consider only multimorbid patients, while **Figure 4** exhibits a barplot of the age ([0-49] years, [50-74] years and [>75] years) prevalence among different malignant neoplasms of stomach, consider only multimorbid patients.

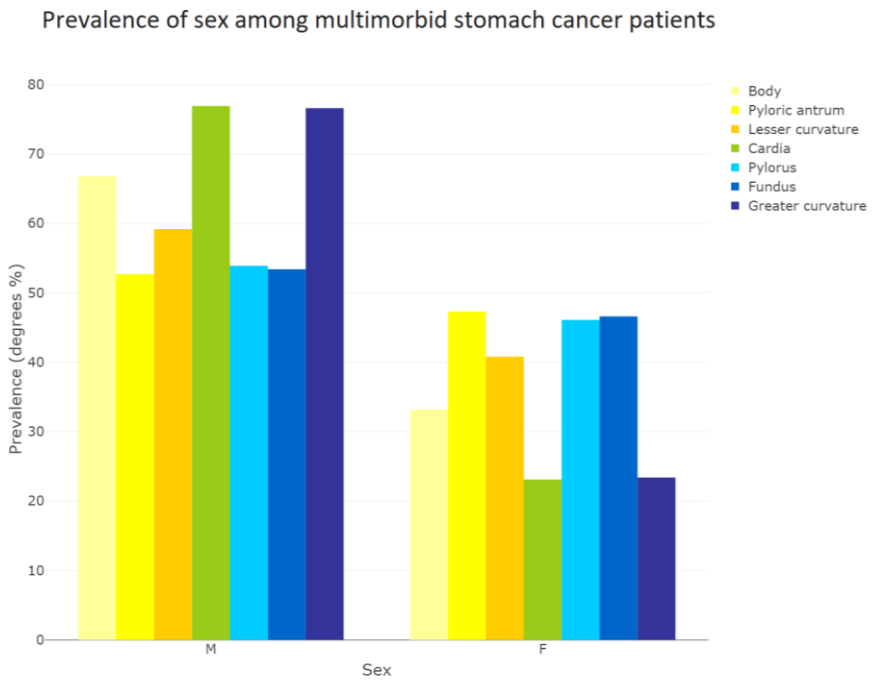

Figure 3– Prevalence of sex, Male or Female, for each type of malignant neoplasm of stomach

Looking at the figure above, it should be noted that, in general, any type of stomach cancer affected males more than females, within the Portuguese population, between 2011 and 2015. Particularly, the most prevalent malignant neoplasms of the stomach in Portugal, for the male sex, were malignant neoplasms of the cardia of the stomach (76.9%), accompanied by the greater curvature (76.6%) and the body of the stomach (66.8%). In contrast, for women patients, the most prevalent malignant neoplasms of the stomach, were malignant neoplasms of the pyloric antrum of the stomach (47.3%), accompanied by the fundus of the stomach (46.6 %) and the pylorus of the stomach (46.1%).

From the analysis of Figure 4, it can be highlighted that, between 2011 and 2015, for ages between 50 and 74 years, the most prevalent malignant neoplasms of the stomach in Portugal were the malignant neoplasm of the body of the stomach (56.2%), followed by the lesser curvature (54.7%) and the greater curvature of the stomach (51.6%). In addition, for ages over 75 years, the most prevalent malignant neoplasms of the stomach were the malignant neoplasm of the pyloric antrum of the stomach (62.5%), followed by fundus (61.1%) and the pylorus (58.0%).

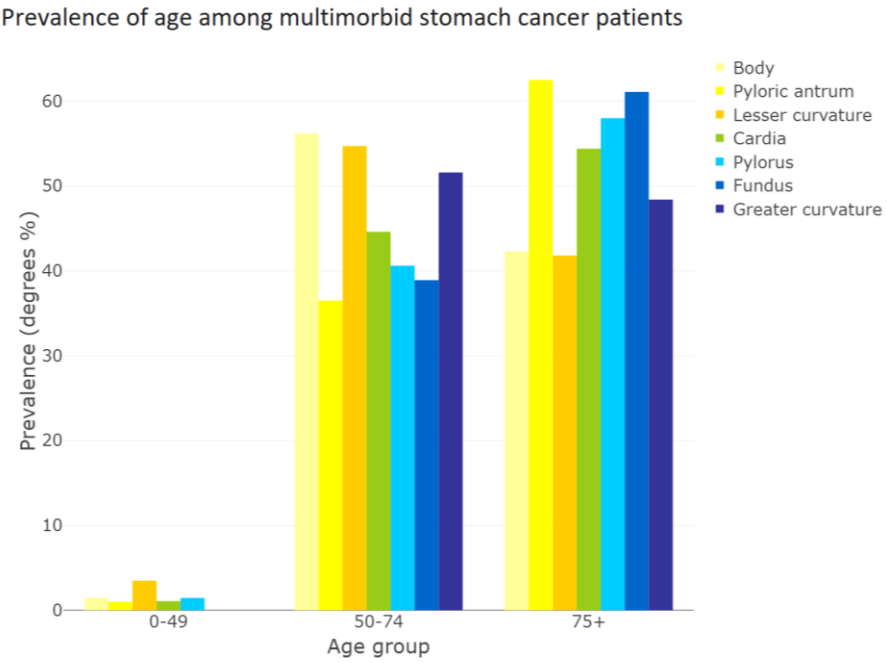

Figure 4– Prevalence of age, [0-49] years, [50-74] years and [>75] years for each type of malignant neoplasm of stomach

3.4. Clustering

3.4.1. Optimal number of clusters and internal validation

For the Elbow method, the optimal number of clusters was identified by selecting the k value after which the WSS score does not significantly decrease. This is called the inflection point. However, recognize that point depends on manually viewing, and are considered an ambiguous approach. Most of the representation follows a similar pattern, with a graph with alternating peaks, as **Figure 5** reveals.

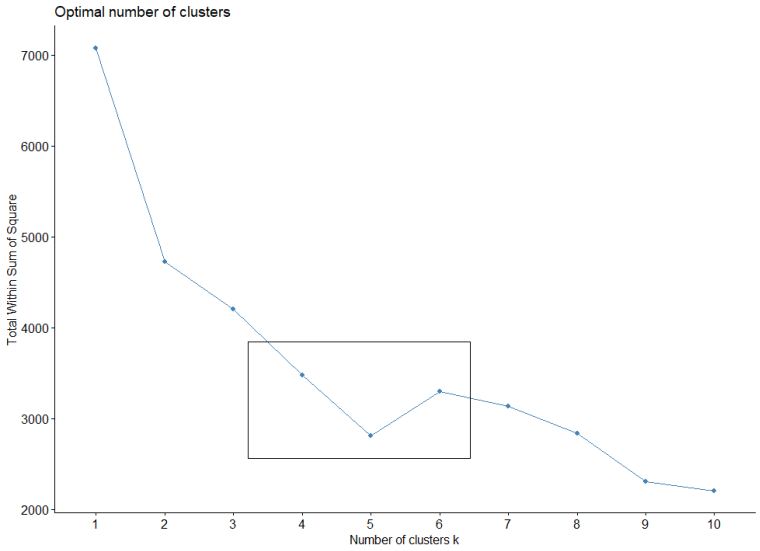

Figure 5 - Elbow method representation of the number k of clusters as a function of the sum of squared distances, with no discrimination elbow point identification in the squared area.

The next step was to calculate the silhouette score through the functions **clust.stats ()** and **clValid (),** which allowed selecting of the optimal number of clusters and the most suitable algorithm for the stomach cancer type and anatomical groups database (**Table 4**).

Table 7 – Results of optimal number of clusters and clustering algorithm for each malignant neoplasm of stomach cancer and anatomic groups, by using cluster.stats() and clValid() functions

Type of stomach cancer	Algorithm	Clust.stats()		clValid()
		Optimal number of clusters (Silhouette score)	**Avg.silh.width**	Silhouette score
Malignant neoplasm of body of stomach	k-means	2	**0,26**	0,27
	PAM	2	0,24	0,24
	Hierarchical	3	0,24	**0,47**
Malignant neoplasm of pyloric antrum	k-means	2	0,26	0,28
	PAM	2	0,26	0,27
	Hierarchical	2	**0,30**	**0,33**
Malignant neoplasm of lesser curvature of stomach, unspecified	k-means	3	**0,26**	0,28
	PAM	2	0,23	0,25
	Hierarchical	4	0,23	**0,30**
Malignant neoplasm of cardia	k-means	3	0,28	0,26
	PAM	4	**0,31**	0,24
	Hierarchical	4	0,26	**0,40**
Malignant neoplasm of pylorus	k-means	4	0,28	**0,40**
	PAM	2	0,26	0,22
	Hierarchical	2	**0,38**	0,39
Malignant neoplasm of fundus of stomach	k-means	2	0,25	**0,29**
	PAM	2	**0,28**	0,25
	Hierarchical	2	0,25	0,27
Malignant neoplasm of greater curvature of stomach, unspecified	k-means	4	0,28	0,28
	PAM	3	**0,29**	0,23
	Hierarchical	2	0,28	**0,29**
Upper Stomach	k-means	3	0.27	0,25
	PAM	3	0,26	0,26
	Hierarchical	2	**0,29**	**0,30**
Lower Stomach	k-means	3	0,28	0,28
	PAM	2	0,26	0,27
	Hierarchical	2	**0,28**	**0,33**

Attempting to the Table 6, the scores obtained for both functions differed for the majority of stomach cancer types. In this sense, the highest score for the silhouette measure, always obtained with **clValid(),** was used to select the best clustering algorithm that would be applied to obtain disease groups for the different types of stomach cancer. Globally, each dataframe will submitted to hierarchical grouping, except for the malignant neoplasm of the Pylorus and Fundus stomach cancer, where the k-means method will be applied. The optimal number of clusters for **clValid()** function was always two, that´s why is not present in the table.

3.4.2. Clinically Meaningful Multimorbidity Clusters

One of the main purposes of the present work is interpreting the clusters obtain for each type of stomach cancer.

Table 8 – Meaningful groupings obtained for each type of stomach cancer and anatomical groups, with reference to the most prevalent chronic diseases in each one

Type of stomach cancer	Cluster 1	Cluster 2
Malignant neoplasm of body of stomach	(N = 169) **DIAB_UC** 152 (90%) **CHF** 60 (36%) **COPD** 41 (24%)	(N = 123) **COPD** 61 (50%) **MILDLD** 36 (29%) **RD** 35 (28%)
Malignant neoplasm of pyloric antrum	(N = 344) **CHF** 157 (46%) **COPD** 136 (40%) **RD** 101 (29%)	(N = 325) **DIAB_UC** 325 (100%) **COPD** 82 (25%) **CHF** 76 (23%)
Malignant neoplasm of lesser curvature of stomach, unspecified	(N = 95) **CHF** 47 (49%) **COPD** 37 (39%) **DIAB_UC** 31 (33%)	(N = 54) **DIAB_UC** 54 (100%) **COPD** 13 (24%) **MILDLD** 11 (20%)
Malignant neoplasm of cardia	(N = 143) **COPD** 59 (41%) **RD** 51 (36%) **CEVD** 34 (24%)	(N = 102) **DIAB_UC** 102 (100%) **COPD** 21 (21%) **CHF** 16 (16%); **MI** 16 (16%)
Malignant neoplasm of pylorus	(N = 19) **COPD** 11 (58%) **CHF** 10 (53%) **MILDLD** 7 (37%)	(N = 18) **DIAB_UC** 16 (89%) **CHF** 5 (28%) **COPD** 5 (28%)
Malignant neoplasm of fundus of stomach	(N = 30) **DIAB_UC** 30 (100%) **RD** 8 (27%) **CHF** 7 (23%)	(N = 25) **COPD** 13 (52%) **CHF** 11 (44%) **CEVD** 7 (28%)
Malignant neoplasm of greater curvature of stomach, unspecified	(N = 30) **DIAB_UC** 24 (80%) **CEVD** 14 (47%) **CHF** 8 (27%)	(N = 10) **RD** 5 (50%); **COPD** 5 (50%) **CHF** 3 (30%); **MILDLD** 3 (30%); **CANCER** 3 (30%)
Upper Stomach	(N = 312) **DIAB_UC** 312 (100%) **COPD** 80 (26%) **CHF** 75 (24%)	(N = 280) **COPD** 120 (43%) **CHF** 94 (34%) **RD** 89 (32%)
Lower Stomach	(N = 363) **CHF** 168 (46%) **COPD** 145 (40%) **RD** 103 (28%)	(N = 342) **DIAB_UC** 342 (100%) **COPD** 89 (26%) **CHF** 80 (23%)

Of all the clusters obtained, the most frequent group are DIAB_UC, CHF, and COPD. However, there are an exception to malignant neoplasm of fundus of stomach and malignant neoplasm of greater curvature of stomach, unspecified, in which CHF and COPD appears with CEVD. Another observation is that one of the diseases above are also present in the second cluster. DIAB_UC only appears in the two clusters of malignant neoplasm of greater curvature of stomach, unspecified. In majority, CHF makes part of the two clusters, with the exception for malignant neoplasm of body of stomach and malignant neoplasm of lesser curvature of stomach, unspecified. COPD are always present in two clusters, with the exception for malignant neoplasm of fundus of stomach and malignant neoplasm of greater curvature of stomach, unspecified. Finally, in general, RD appears as a prevalent chronic disease in one of the two clusters for each type of stomach cancer, with the exception to malignant neoplasm of lesser curvature of stomach, unspecified and pylorus.

After a median calculation for number of comorbidities for each cluster for each type of stomach cancer, the most patients with multimorbidity presents two chronic diseases.

3.5. Association Rules

This section presents a table with the highest support rules for each malignant neoplasm of stomach, as the assumption is to find relevant diseases rules to a characterization of multimorbid stomach cancer patients' patterns. The minimum support was chosen to 2%, and the minimum confidence was set at 50%. Moreover, a cut of six rules was considered, because for a greater number of rules, the support value will be increasingly reduced and, therefore, they lose relevance. The support and confidence measures were presented as percentage.

Table 9 – Top association rules of chronic Charlson' comorbidities for each malignant neoplasms of stomach, with a cut off of six rules "(…)"

	Association rules	**Support (%)**	**Confidence (%)**	**Lift**	**Count**
Malignant neoplasm of body of stomach	[1] {CHF} => {DIAB_UC} [2] {CEVD} => {DIAB_UC} [3] {MSLD} => {MILDLD} [4] {DIAB_C} => {RD} [5] {DEM} => {DIAB_UC} [6] {MI, CHF} => {DIAB_UC}	[1] 15,07 [2] 8,22 [3] 5,48 [4] 5,48 [5] 3,08 [6] 2,74	[1] 52,38 [2] 53,33 [3] 88,89 [4] 66,67 [5] 52,94 [6] 50,00	[1] 0.944 [2] 0.961 [3] 3.762 [4] 3.673 [5] 0.954 [6] 0.901	[1] 44 [2] 24 [3] 16 [4] 16 [5] 9 [6] 8
Malignant neoplasm of pyloric antrum	[1] {DIAB_C} => {RD} [2] {MSLD} => {MILDLD}	[1] 4,04 [2] 2,84	[1] 50,00 [2] 67,86	[1] 2.438 [2] 4.011	[1]27 [2]19
Malignant neoplasm of lesser curvature of stomach, unspecified	[1] {COPD} => {DIAB_UC} [2] {RD} => {CHF} [3] {MILDLD} => {DIAB_UC} [4] {MI} => {DIAB_UC} [5] {DIAB_UC, RD} => {CHF} [6] {CHF, RD} => {DIAB_UC} (…)	[1] 16,78 [2] 11,41 [3] 10,07 [4] 6,71 [5] 6,04 [6] 6,04	[1] 50,00 [2] 51,51 [3] 57,69 [4]55,56 [5] 56,25 [6] 52,94	[1] 0.876 [2] 1.633 [3] 1.011 [4] 0.974 [5] 1.783 [6] 0.928	[1]25 [2]17 [3]15 [4]10 [5]9 [6]9
Malignant neoplasm of cardia	[1] {CEVD} => {DIAB_UC} [2] {MI} => {DIAB_UC} [3] {DIAB_C} => {RD} [4] {MSLD}=> {MILDLD} [5] {MI, CHF} => {RD}	[1] 9,80 [2] 9,39 [3] 7,35 [4] 2,45 [5] 2,04	[1] 51,06 [2] 56,10 [3] 60,00 [4] 66,67 [5] 50,00	[1] 1.034 [2] 1.136 [3]2.262 [4] 4.537 [5] 1.885	[1]24 [2]23 [3]18 [4]6 [5]5
Malignant neoplasm of pylorus	[1] {MILDLD} => {CHF} [2] {RD} => {DIAB_UC} [3] {CANCER} => {COPD} [4] {PVD} => {CEVD}	[1] 13,51 [2] 13,51 [3] 8,11 [4] 8,11	[1] 71,43 [2] 62,50 [3] 60,00 [4] 60,00	[1] 1.762 [2]1.360 [3]1.388 [4] 3.700	[1]5 [2]5 [3]3 [4]3

	[5] {PVD} => {COPD} [6] {CEVD}=> {PVD} (…)	[5] 8,11 [6] 8,11	[5] 60,00 [6] 50,00	[5] 1.388 [6] 3.700	[5]3 [6]3
Malignant neoplasm of fundus of stomach	[1] {RD} => {DIAB_UC} [2] {PVD}=> {DIAB_UC} [3] {MI} => {CHF} [4] {MI} => {DIAB_UC} [5] {MI, CHF} => {DIAB_UC} [6] {MI, DIAB_UC} => {CHF} (…)	[1] 14,55 [2] 10,91 [3] 9,09 [4] 9,09 [5] 7,27 [6] 7,27	[1] 66,67 [2] 75,00 [3] 62,50 [4] 62,50 [5] 80,00 [6] 80,00	[1] 1.078 [2] 1.213 [3] 1.910 [4]1.011 [5] 1.294 [6] 2.444	[1]8 [2]6 [3]5 [4]5 [5]4 [6]4
Malignant neoplasm of greater curvature of stomach, unspecified	[1] {CEVD} => {DIAB_UC} [2] {CHF} => {DIAB_UC} [3] {PARA} => {CEVD} [4] {DEM} => {DIAB_UC} [5] {MI}=> {CHF} [6] {PUD} => {DIAB_UC} (…)	[1] 20,00 [2] 17,50 [3] 5,00 [4]5,00 [5] 5,00 [6] 5,00	[1] 57,14 [2] 63,64 [3] 100,00 [4]100,00 [5] 100,00 [6] 66,67	[1] 0.914 [2] 1.018 [3] 2.857 [4] 1.600 [5] 3.636 [6] 1.067	[1]8 [2]7 [3]2 [4]2 [5]2 [6]2
Upper stomach	[1] {CEVD} => {DIAB_UC} [2] {MI} => {DIAB_UC} [3] {DIAB_C} => {RD} [4] {MSLD} => {MILDLD} [5] {DEM} => {DIAB_UC}	[1] 8,78 [2] 8,11 [3] 5,74 [4] 4,05 [5] 2,53	[1] 52,00 [2] 53,33 [3] 60,71 [4] 82,76 [5] 60,00	[1] 0.971 [2] 0.996 [3] 2.765 [4] 4.374 [5] 1.121	[1]52 [2]48 [3]34 [4]24 [5]15
Lower stomach	[1] {DIAB_C} => {RD} [2] {MSLD} => {MILDLD}	[1] 3,83 [2] 2,84	[1] 50,00 [2] 68,97	[1] 2.431 [2] 4.052	[1]27 [2]20

The following lines describe the characterization of association rules obtained among diseases, for each cancer type of stomach cancer and anatomical groups.

- **Malignant neoplasm of body of stomach** admits 5 dyads and 1 triad. DIAB_UC is present in five of the six most relevant rules. The higher support occurs when diabetes without complications is linked to CHF and CEVD, respectively.
- For **Malignant neoplasm of pyloric antrum**, there are only two dyads of comorbidities, which are Diabetes with complications related to Moderate-to-severe renal disease and Mild liver disease with Moderate or severe liver disease
- **Malignant neoplasm of lesser curvature of stomach, unspecified,** stomach admits 4 dyads and 2 triad. The two top rules are {COPD} => {DIAB_UC} and {RD} => {CHF}. Relative to the latter (supp=11,4%), when in the presence of DIAB_UC, support declines to 6%.
- **Malignant neoplasm of cardia** presents 4 dyads and 1 triad. Higher dyad support is observed for DIAB_UC linked to related to CEVD (supp=9,80%) and MI (supp=9,39%)
- For **Malignant neoplasm of pylorus** there are six dyads of morbidities. PVD is present in three of six; COPD and CEVD are present in two of six. The strongest rules include {MILDLD} => {CHF} and {RD} => {DIAB_UC}. Particularly, the rule 1 {MILDLD} => {CHF} presents, simultaneously, the highest support and confidence.
- **Malignant neoplasm of fundus of stomach** admits 4 dyads and 2 triad. Of the 6 most supported rules, DIAB_UC appears in 5 of them and is present in the first two, related to
- RD and PVD. Also, is more probable to occur dyads(supp=9%) {MI} => {CHF} and {MI} => {DIAB_UC}, in comparison to correspondent triad(supp=7%), {MI, CHF} => {DIAB_UC}
- For **Malignant neoplasm of greater curvature of stomach, unspecified,** there are six dyads of comorbidities. Of the 6 most supported rules, DIAB_UC appears in 4 of them and is present in the first two, related to CEVD and CHF, in a descent support.

- A comparison between relationships of malignant neoplasm of **cardia,fundus and body** of stomach:

    a) The cardia and fundus location tumor share {CEVD} => {DIAB_UC}, {DIAB_C} => {RD} and {MSLD}=> {MILDLD} relations.
    b) The body and fundus only share the triad {MI, CHF} => {DIAB_UC}.
    c) Cardia and fundus have in common only {MI} => {DIAB_UC}.
    d) Comparing now with the anatomical group "**Upper stomach**", selected at the beginning of the study, there are similarities, as expected, with the conclusions of lines a), b) and c), sharing the following rules: DIAB_UC linked to CEVD and MI, DIAB_C with RD and MSLD with MILDLD.

- Malignant neoplasm of **pyloric antrum and pyloru**s don´t share any relationships between chronic diseases.
    a) The anatomical group "**Lower stomach**" presents the same rules of malignant neoplasm of pyloric antrum

3.4. Network Analysis

Network analysis was applied to demonstrate graphically the complicated nature of interactions between comorbidities, so that, inspect patterns of diseases in patients with multimorbidity. The edge thickness is proportional to the correlation between each disease pair.

1- Malignant neoplasm of the body of the stomach

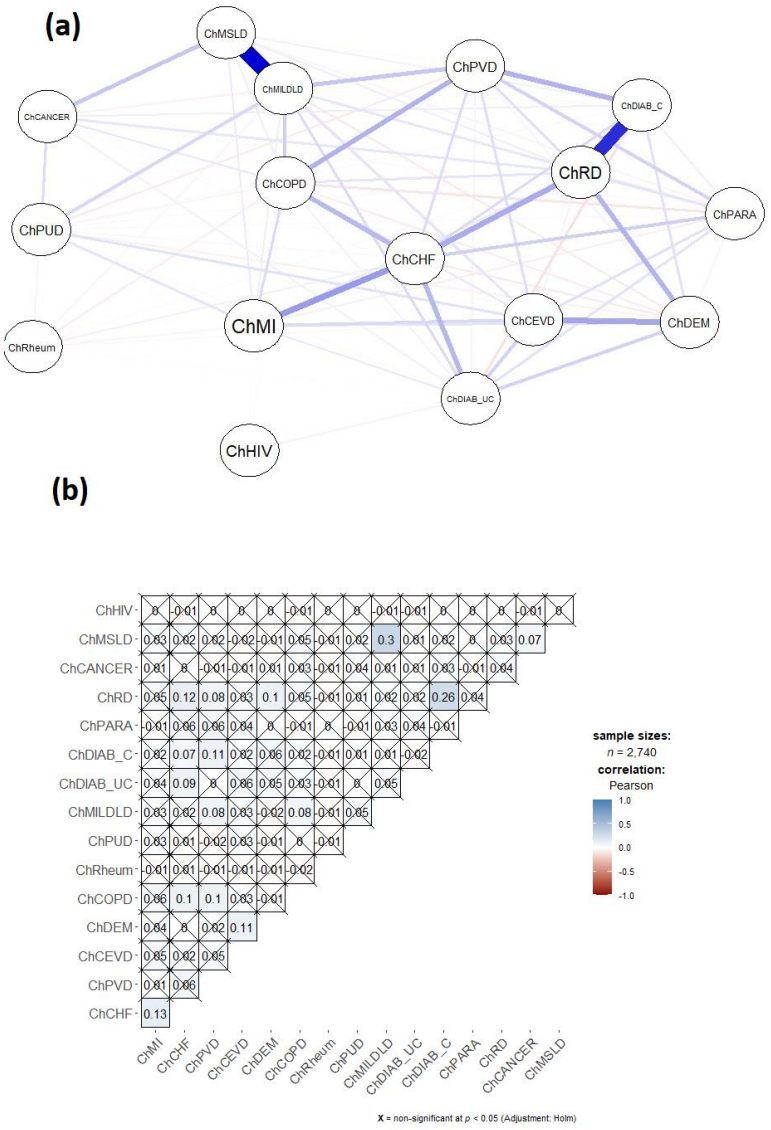

Figure 6– (a) Network representation for malignant neoplasm of comorbities of Charlson for patients with body' stomach; (b)correlogram with Pearson correlation values for pairs of chronic Charlson' comorbidities

Across all malignant neoplasm of body of stomach patients, the most commonly pairs of diseases are Moderate or severe liver disease (MSLD) with Mild liver disease (MILDLD), followed by Moderate-to-severe renal disease (RD) with Diabetes mellitus with chronic complications (DIAB_C). This result can be confirmed by thickness visualization in (**Figure7a**), and by correlogram calculating (**Figure7b**) shown by a cross on the correlation coefficients, reveals the high correlation (0.3 and 0.26, respectively) of these two pair. the same reasoning will be applied in the conclusions of the next diagrams.

2 - Malignant neoplasm of the pyloric antrum of the stomach

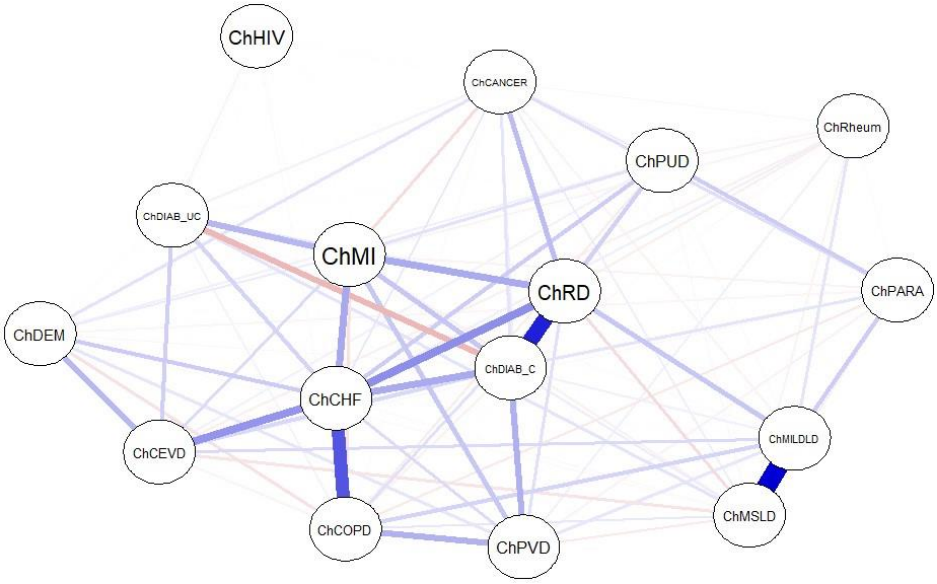

Figure 7 - Network representation for Malignant neoplasm of the pyloric antrum of the stomach

3 - Malignant neoplasm of the lesser curvature of stomach, unspecified

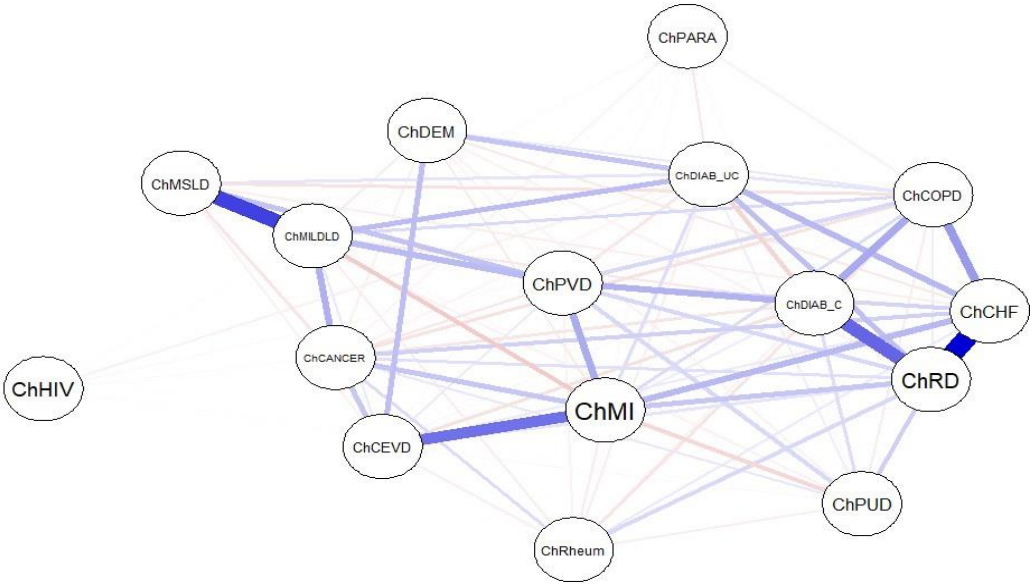

Figure 8- Network representation for Malignant neoplasm of the lesser curvature of stomach, unspecified

4 - Malignant neoplasm of the cardia of the stomach

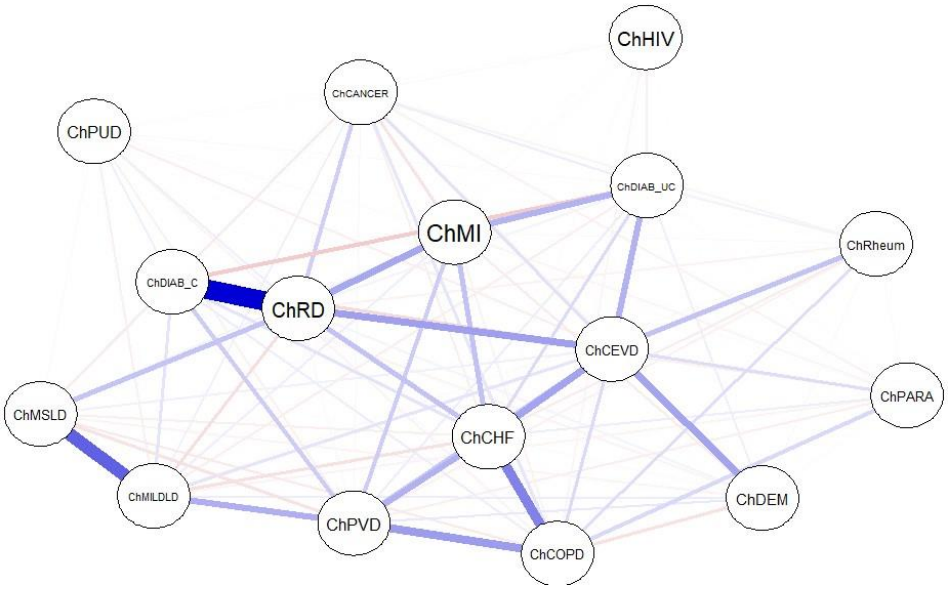

Figure 9 - Network representation for Malignant neoplasm of the cardia of the stomach

5 - Malignant neoplasm of the pylorus of the stomach

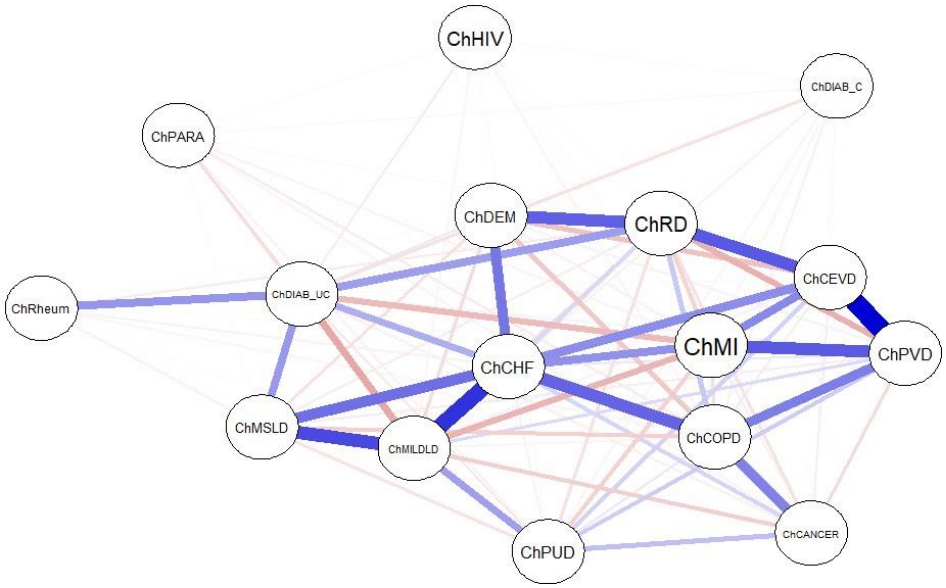

Figure 10 - Network representation for Malignant neoplasm of the pylorus of the stomach

6 - Malignant neoplasm of the fundus of the stomach

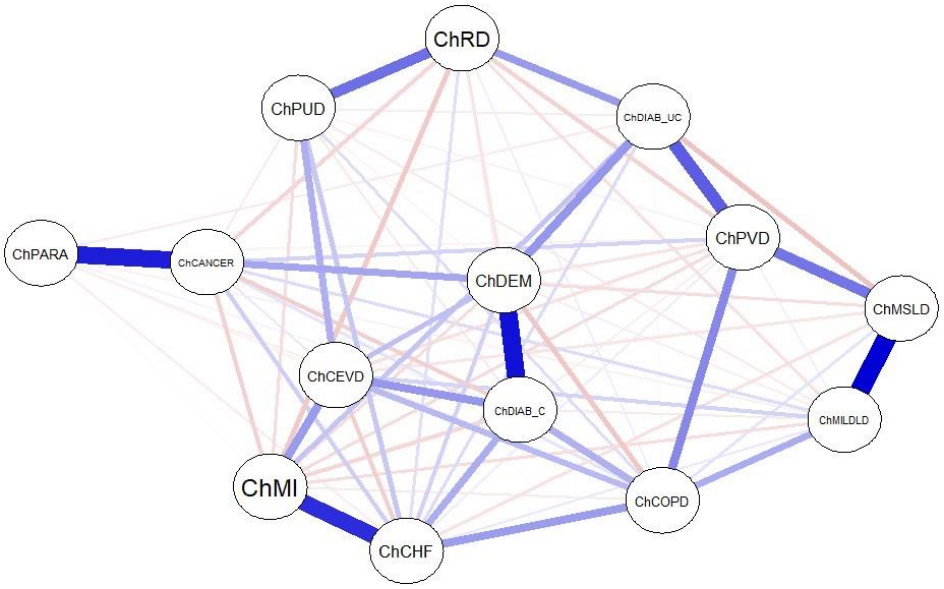

Figure 11 - Network representation for Malignant neoplasm of the fundus of the stomach

7 - Malignant neoplasm of the greater curvature of stomach, unspecified

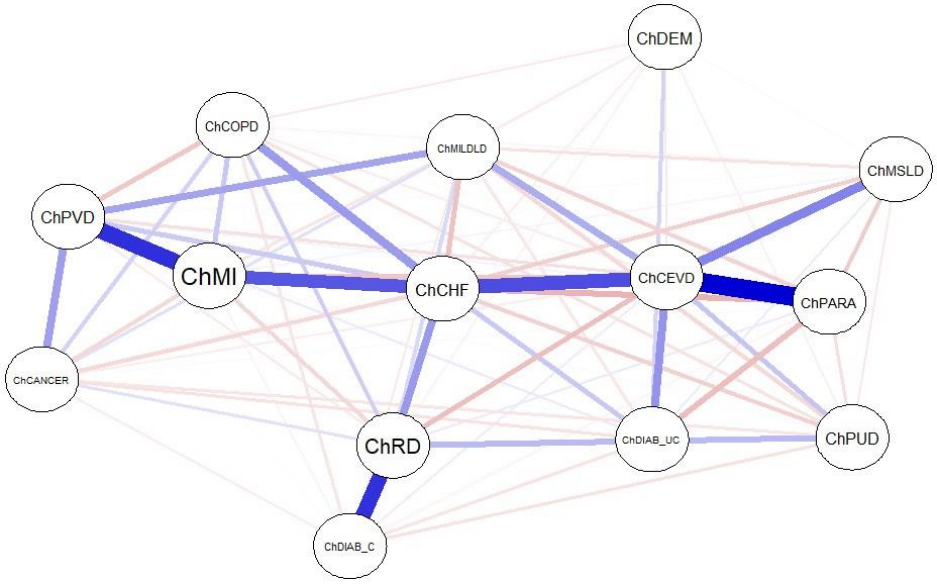

Figure 12- Network representation for Malignant neoplasm of the greater curvature of stomach, unspecified

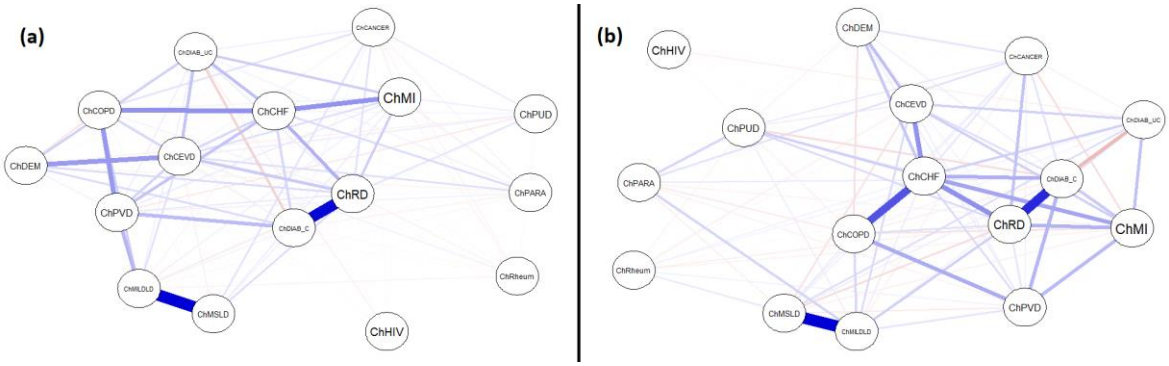

Figure 13 - Network representation for anatomical groups (a)Upper stomach (b) Lower stomach

For **Malignant neoplasm of lesser curvature of stomach, unspecified**, **Malignant neoplasm of pyloric antrum**, and anatomic groups (**Upper and lower stomach**) most correlated pairs included 1) MSLD with MILDLD and 2) RD with DIAB_C.

For **Malignant neoplasm of cardia,** the more weighted pair was RD with CHF, followed by MSLD with MILDLD.

From Malignant neoplasm of pylorus onwards, the visualization of the most correlated pairs was more difficult, due to the high number. In a descent relevance, the CEVD-PVD pair was more visible for **Malignant neoplasm of pylorus** patients, followed by MILDLD-CHF, PUD-MI, and MSLD-MILDLD. **Malignant neoplasm of fundus of stomach** presents MSLD-MILDLD as the most correlated pair, after DEM-DIAB_C, and finally CANCER-PARA. **Malignant neoplasm of greater curvature of stomach, unspecified** network demonstrated a strong correlation between PARA and CEVD, PVD and MI, and lastly, RD with DIAB_C, in a descent order.

Particularly, for **Upper stomach group**, MSLD with MILDLD and RD with DIAB_C, were the most common pairs, with the same correlation, while **Lower stomach**, presents the same pair plus COPD with CHF

A global conclusion admits the most present pairs were MSLD with MILDLD and RD with DIAB_C.

3.5. Generalized Mixed Linear models to analyse Mortality and LOS

Firstly, a table with the results of the mortality models are presented (**Table 10**). Models' coefficients are expressed as odds ratios, along with their respective confidence intervals estimates for the different controls included, namely patient demographics (age, sex, geographic region) and clinical characteristics (presence of metastatic tumor and multimorbidity). The outcome contains the three models previously clarified in sub-section 2.9. Regarding models' goodness of fit measures, marginal/conditional R2 were 0.12/0.14, 0.19/0.14 and 0.13/0.15 for models 1, 2 and 3, respectively, meaning that the inclusion of hospital as random effect did not affect the amount of variation explained by the model. Akaike information criterion (AIC) values, which computes the amount of information lost by a given model, were lower for Model 3 (AIC = 4195.771), followed by Model 2 (AIC = 4537.189), and Model 1 (AIC = 8723.569).

Table 10 – Odds ratio and confidence interval of each predictor for mortality outcome

Death	*Cardia vs Pylorus*		*Cardia/Fundus/Body only*		*Pyloric/antrum only*	
Predictors	*Odds Ratios (OR)*	*CI*	*Odds Ratios (OR)*	*CI*	*Odds Ratios (OR)*	*CI*
(Intercept)	0.07**	0.05 – 0.10	0.06**	0.04 – 0.11	0.05**	0.03 – 0.09
SEX [Male]	1.21*	1.08 – 1.35	1.20*	1.02 – 1.42	1.21*	1.04 – 1.42
FaixaEtaria50-74	1.17*	0.91 – 1.50	1.01*	0.73 – 1.40	1.51*	1.00 – 2.28
FaixaEtaria [75+]	2.01**	1.56 – 2.60	1.69*	1.21 – 2.35	2.67**	1.76 – 4.03
Multimorb level [1 comorbidity]	1.29*	1.08 – 1.52	1.19*	0.93 – 1.53	1.35*	1.07 – 1.71
Multimorb level [2 comorbidities]	1.57**	1.29 – 1.92	1.47*	1.10 – 1.97	1.63**	1.25 – 2.14
Multimorb level [3 comorbidities]	2.02**	1.54 – 2.65	1.57*	1.05 – 2.35	2.49**	1.72 – 3.60
Multimorb level [4+ comorbidities]	2.45**	1.69 – 3.57	1.62*	0.95 – 2.77	3.63**	2.14 – 6.16
Mets [1 – With Metastasis]	2.78*	2.42 – 3.19	2.95**	2.41 – 3.62	2.64**	2.18 – 3.20
Type [pylorus/pyloric antrum]	0.88*	0.79 – 0.98				
Region [Algarve]	1.12*	0.65 – 1.93	1.49*	0.77 – 2.89	0.81*	0.38 – 1.74
Region [Área Metropolitana de Lisboa]	0.96*	0.70 – 1.31	1.26*	0.82 – 1.92	0.74*	0.49 – 1.12
Region [Centro]	0.84*	0.61 – 1.16	1.05*	0.68 – 1.61	0.76*	0.50 – 1.15
Region [Norte]	0.88*	0.63 – 1.21	1.02*	0.67 – 1.55	0.76*	0.51 – 1.15

Reference categories: Sex→Female; FaixaEtaria→ [0-50] years; Multimorb level→ 0 comorbidity; Mets→ 0 – Without Metastasis; Type →cardia/body/fundus; Region→Alentejo

When considering model 1, in which it is possible to compare the effects between cardia and pylorus (each made part of upper and lower region of stomach, respectively) the pylorus/pyloric antrum cancer admits less death in comparison to cardia region (OR 0.88, 95% CI 0.79 – 0.98). The regions of Centro (OR 0.84, 95% CI 0.61 – 1.16), Norte (OR 0.88, 95% CI 0.63 – 1.21) and Área Metropolitana de Lisboa(OR 0.96, 95% CI 0.70 – 1.31) are positively related to death in comparison with reference category (Alentejo region). In contrast, solid metastatic tumor (OR 2.78, 95% CI 2.42 – 3.19) and the presence of four or more comorbidities (OR 2.45, 95% CI 1.69 – 3.57) are negatively related to death, in which an increase above 100% in the odds of death is expected relatively to the reference categories (without metastasis and without comorbidities).

Regarding model 2, in which the focus was only on the upper region of stomach (Cardia/Fundus/Body only), all predictors were negatively related with death. Particularly, ages higher than 75 years old (OR 1.69, 95% CI 1.21 – 2.35), presence of metastatic solid tumor (OR 2.95, 95% CI 2.41 – 3.62) and the presence of four or more comorbidities (OR 1.62, 95% CI 0.95 – 2.77) were the most independent variables with higher odds of death rate.

For last, considering model 3, which focused on the lower region of stomach (Pyloric/antrum only), regions Área Metropolitana de Lisboa (OR 0.74, 95% CI 0.49 – 1.12), Centro (OR 0.76, 95% CI 0.50 – 1.15), Norte (OR 0.76, 95% CI 0.51 – 1.15) and Algarve (OR 0.81, 95% CI 0.38 – 1.74) are also positively related to death in comparison with reference category (Alentejo region). Like the previous ones, the presence of four or more comorbidities (OR 3.63, 95% CI 2.14 – 6.16) and metastatic solid tumor (OR 2.64, 95% CI 2.18 – 3.20) substantially increased the odds of death.

Length of stay

For assessing length of stay (LOS), variables representing patient demographics (age, sex, region) and clinical characteristics (presence of metastatic tumor and multimorbidity) also constituted the predictors for LOS of stomach cancer hospitalizations (See **Table 11**). The outcome contains the three models previously clarified in subsection 2.9. Regarding models' goodness of fit measures, marginal/conditional R2 were 0.11/0.41, 0.13/0.44 and 0.15/0.53 for models 1, 2 and 3, respectively, meaning that the inclusion of hospital as random effect more than doubled the amount of variation explained by the model. Akaike information criterion (AIC) values, which computes the amount of information lost by a given model, were lower for Model 2 (AIC = 37858.245), followed by Model 3 (AIC = 38632.329), and Model 1 (AIC = 77339.272).

Table 11 - Incidence rate ratios and confidence interval of each predictor for hospital' length of stay outcome

Length of stay(LOS)	*Cardia vs Pylorus*		*Cardia/Fundus/Body only*		*Pyloric/antrum only*	
Predictors	*Incidence Rate Ratios (IRR)*	*CI*	*Incidence Rate Ratios (IRR)*	*CI*	*Incidence Rate Ratios (IRR)*	*CI*
(Intercept)	8.97**	8.26 – 9.73	10.30**	9.32 – 11.39	8.37**	7.50 – 9.35
SEX [Male]	1.03**	1.02 – 1.04	1.00*	0.98 – 1.02	1.04**	1.03 – 1.06
FaixaEtaria50-74	1.12**	1.09 – 1.15	1.13**	1.09 – 1.18	1.10**	1.06 – 1.13
FaixaEtaria [75+]	1.14**	1.12 – 1.17	1.11**	1.07 – 1.16	1.17**	1.13 – 1.21

Multimorb level [1 comorbidity]	1.04**	1.03 – 1.06	1.04*	1.02 – 1.07	1.05**	1.03 – 1.07
Multimorb level [2 comorbidities]	1.23**	1.21 – 1.25	1.36**	1.32 – 1.40	1.15**	1.12 – 1.18
Multimorb level [3 comorbidities]	1.31**	1.27 – 1.35	1.38**	1.32 – 1.44	1.25**	1.20 – 1.30
Multimorb level [4+ comorbidities]	1.30**	1.24 – 1.36	1.35**	1.26 – 1.44	1.25**	1.17 – 1.34
Mets [1 – With Metastasis]	1.03*	1.01 – 1.04	0.95**	0.93 – 0.97	1.10**	1.08 – 1.12
Type [pylorus/pyloric antrum]	1.07**	1.06 – 1.08				
Region [Algarve]	1.13*	1.03 – 1.24	1.19*	1.05 – 1.34	1.00*	0.38 – 1.74
Region [Área Metropolitana de Lisboa]	1.28**	1.21 – 1.35	1.14*	1.06 – 1.23	1.45**	0.49 – 1.12
Region [Centro]	1.21**	1.13 – 1.28	1.08*	0.99 – 1.17	1.35**	0.50 – 1.15
Region [Norte]	1.21**	1.12 – 1.30	0.97*	0.88 – 1.07	1.49**	0.51 – 1.15

Reference categories: Sex→Female; FaixaEtaria→[0-50]years; Multimorb level→ 0 comorbidity; Mets→ 0 – Without Metastasis; Type → cardia/body/fundus; Region→ Alentejo

The analysis of LOS attempting to Incidence Rate Ratios and results differed from the death' analysis. In model 1, all predictors were negatively related with LOS, with significantly higher LOS to Área Metropolitana de Lisboa (IRR 1.28, 95% CI 1.21 – 1.35) in comparison with the reference category (Alentejo region), three (IRR 1.31, 95% CI 1.27 – 1.35) and four comorbidities (IRR 1.30, 95% CI 1.24 – 1.36) and for pylorus/pyloric antrum cancer (IRR 1.07, 95% CI 1.06 – 1.08).

Regarding model 2, focusing only on the upper region of stomach (Cardia/Fundus/Body only), metastatic solid tumor (IRR 0.95, 95% CI 0.93 – 0.97) and Norte region (IRR 0.97, 95% CI 0.88 – 1.07) are positively related to LOS. In contrast, the presence of three (IRR 1.38, 95% CI 1.32 – 1.44) and two comorbidities (IRR 1.36, 95% CI 1.32 – 1.40) are the most associated with longer hospitalizations.

For last, model 3 focused on the lower region of stomach (Pyloric/antrum only), showing that all predictors were negatively related with LOS, particularly in the Norte region (IRR 1.49, 95% CI 0.51 – 1.15), Área Metropolitana de Lisboa (IRR 1.45, 95% CI 0.49 – 1.12), Centro (IRR 1.35, 95% CI 0.50 – 1.15), and when considering the presence of three (IRR 1.25, 95% CI 1.20 – 1.30), and four comorbidities (IRR 1.25, 95% CI 1.17 – 1.34).

ICC estimation was 0.33, 0.36 and 0.45 for models 1, 2 and 3, respectively, meaning that the proportion of the variance explained by the fact that episodes occur in different hospitals ranged from 33 to 45%, with the highest variation found for the subsample composed by those affected by cancer in the lower region of the stomach (model 3).

4.Discussion of results

4.1. Relationship with Existing Literature

The present findings suggest a higher prevalence of malignant neoplasm of cancer, in Portugal, between 2011 and 2015, for men compared to women, considering both for non-multimorbidity and multimorbidity conditions. Regarding age, in the absence of multimorbidity, the Portuguese population with stomach cancer was, in the vast majority, aged between 50 and 74 years, while considering multimorbidity hospitalizations, patients were 75 years or older. In 2017, Castro et *al.* [14] published an article that predicted the incidence of stomach cancer in Portugal for 2015, finding a higher incidence in males compared to females. Another similar findings were presented in other studies [7],[11],[100].

Considering only the multimorbidity condition, two main conclusions could be underlined. First, the majority of Portuguese patients with stomach cancer, between 2011 and 2015, presented two chronic diseases and were aged over 75 years. A quite similar conclusion was presented by Gonçalves et *al.*[101], during the year 2015, with an average age for the multimorbid Portuguese population of 59.8 years, being higher in men (62.3 years). Second, the most prevalent malignant neoplasm of the stomach for the Portuguese male was malignant neoplasms of the cardia of the stomach, and for women was malignant neoplasms of the pyloric antrum of the stomach. No Portuguese reports were found, but Crew et *al.* [15] stated in 2006, that Gastric cardia tumors accounted, for nearly half of all stomach cancers among men from US and UK.

Related to clustering algorithms results, the main insight was the consistent clusters comprising diabetes without complications, CHF and COPD. It is possible to find similarities in the results of association rules and networking approaches. For all types of stomach cancer, with the exception of malignant neoplasm of the pyloric antrum and malignant neoplasm of the pylorus, the highest support rule always presented diabetes without complications as a consequent itemset disease.

The first explanation may be related to the relative frequency of these three chronic diseases in the original database. Attempt to section 3.1., the high proportion of diabetes is noticeable (16.5%), followed by COPD (7.21%) and CHF (5.30%). Based on the literature, in 2011, Legler et *al.* [102] concluded that frequently conditions identified as co-occurring with cancer were COPD, diabetes, CVD, and CHF. Two years ago, a similar conclusion was stated by Fowler et *al.* [59]. From theory, the results of the different methods of characterizing disease patterns may have to do with the sharing of risk factors with stomach cancer, as evidenced in[103],[104], in general by unhealthy lifestyle habits.

Particularly, Tseng *et al.* [105] evidenced links between diabetes and gastric cancer, due to the shared risk factors including obesity, insulin resistance, hyperinsulinemia, and smoking. Dunlay et *al.* [106] determined more frequently risk factors for heart failure. Hypertension, obesity, and smoking were more recurrent, and, according to previous remarks, are a share risk factors with stomach cancer. Also, the same author stated that the risk of heart failure was particularly high for coronary disease and diabetes. Again, in agreement with the results obtained, Osman et *al.* [107]found the risk factors most associated with COPD. However, following the study conducted by Mahmoodi et *al*. [108], results of association rules differ from those obtained in the present work, stating that cardiovascular patients are less susceptible to stomach cancer. Specially to Portugal, Castro et *al.* [109] verified that between 1994 to 2009 , dietary habits and smoking are recognised as important gastric cancer determinants

The present study showed sociodemographic and clinical conditions as important factors to assess relevant health outcomes. Assessing multimorbidity and other relevant factors impacting hospital mortality and length of stay contributes with the improvement of decision-making for future patients with stomach cancer. In summary, findings on the assessment of a selected set of specific predictors, including age, sex, area of residence, number of comorbidities (as a proxy of multimorbidity), and presence of metastatic tumor were: 1) higher mortality, in the presence of metastatic tumor, four or more comorbidities, ages between 50 and 74 years and for upper region of the stomach; 2) higher LOS in the presence of metastatic tumor, three comorbidities and ages between 50 to 74 years, and 3) for the lower region, higher LOS in the presence of

metastatic tumor, three and four comorbidities and ages of 75 or more years. Moreover, stomach cancer death is less common to occur in regions Norte and Lisbon, although they presented a higher LOS.

In comparison to upper and lower region of stomach, some insights are relevant, such as the lower rate of mortality and longer length of stay for Pyloric/antrum pyloric patients. An association can be done with Ferronha et *al.* [110], who stated that Cardia stomach cancer, situated in upper region of stomach had a significantly lower survival. The survival rates between 1 and 3 years after diagnosis, varied from 62.3% to 29.0%. In contrast, for the Distal one-third /antrum-pylorus region of stomach, the survival rates ranged from 80.8% to 46.2%, for 1 and 3 years after diagnosis, respectively. Regarding proximal one-third/fundus and middle one-third/body regions survival rates are also lower than antrum-pylorus region in first year of diagnosis, excepted in third year, with a higher rate of survival to Middle one-third/body. In 2015, Morais et *al* [14]had already warned that distal gastric carcinomas cases were the most frequent in Portuguese hospitals.

Conversely, Gonçalves et *al.*[101] analysed events during 2015, from Portuguese National Health Service hospitalisations database, during the year 2015, concluded that worst prognosis was associated with six or more conditions per person.

In this sense, an improved prognosis associated with stomach cancer suggests more intensive prevention and control efforts in Portugal, especially in patients with several comorbidities and in advanced stages of the disease. The relationship of length of stay with the type of stomach cancer should also be taken into account, in the future, to improve hospital management and resources, especially in North and Lisbon regions. In terms of comparison, no study was reported to evaluate the impact of stomach cancer on the length of hospital stay in Portugal. Thereby, more efforts are needed in the future in this area to improve hospital management in Portugal.

4.2. Limitations of the study

Some limitations were detected and should be taken into account in future research. A global challenge in multimorbidity domain is related to the lack of a consensually accepted definition for it, in terms of measurement, population samples, different age ranges and different number and type of comorbidities, and different data sources. Based on diseases codification, Charlson' classification system was applied to 17 binary variables, so it is not generalizable for reporting and analysing other comorbidities, conditions, or population groups. Another situation is the accuracy of the assignment of diagnostic codes by the Portuguese medical coders based on the ICD-9-CM system. The inconsistency between hospitals can interfere with the quality of data, which may eventually explain the discrepancy in the outcomes of this study, for the various regions of Portugal. Also, compare results and performance of clustering algorithms with other studies are not very feasible. In fact, they may differ in 1) cluster structure (cluster shape, size, size difference between cluster and number of clusters), 2) presence of outliers, 3) degree of cluster overlap, and 4) choice of similarity measure. Another issue corresponds to the non-deterministic algorithms applied, such K-means and PAM. For each new iteration running are obtained new different results, even on the same data set. Moreover, it is also important to reinforce the possible underlying quality issues associated with the reuse of healthcare administrative data, whose main purpose is for financing and management, rather than research. Data quality is also strongly linked to the quality of clinical coding (e.g., comprehensive and accurate reporting and coding of comorbidities), and hospitals may substantially differ in terms of clinical coding practices [111].

5. Conclusion

In this dissertation, an exploratory analysis of clinical data and identification of the most frequent representative groups in multimorbidity were performed, using different data mining algorithms such as clustering techniques, association mining, and network analysis. The characterization of co-existing diseases with stomach cancer, and respective patterns over time, could constitute an advance in hospitals' quality-improvement patient-centered care. Moreover, it can be further applied as a guide for clinicians to discover potential health diseases before they become a burden in chronic stomach cancer patients.

Briefly, in the next years, Portugal' health care stakeholders need to focus on presence of two or more comorbidities in stomach cancer patients, specifically on male sex and in advanced stage of cancer (tumor metastasis) as these factors increases the odds of death. Explicitly, diseases such Uncomplicated diabetes, CHF and COPD require special attention in the presence of malignant neoplasm of stomach. Also, is notable a demand for a reinforced hospital' resources management for multimorbid pyloric/antrum stomach cancer patients, which is associated with increased length of stay (a proxy of resource use), and for a better diagnosis, prevention and treatment for multimorbid cardia stomach cancer, which is associated with increased mortality.

The use of Machine Learning Models has been suitable to evaluate multimorbid populations and measure the contributing risk factors to important health outcomes, such as in-hospital mortality and length of stay.

However, there is an insufficiency of studies, not only on multimorbidity but also on the impact of stomach cancer on national health services in Portugal. Furthermore, besides those used methods in the present work, other Machine learning algorithms should be tested in the future to characterize multimorbidity patterns. Finally, another time ahead approach would be to include more comorbidities or long-term health conditions, in addition to Charlson' comorbidities, and other variables such as polypharmacy, frailty and socioeconomic status

6. References

[1] D. Lewin and K. J. Lewin, “Stomach,” in *Modern Surgical Pathology (Second Edition)*, 2009, pp. 673–718.

[2] J. Feher, “The Stomach,” in *Quantitative Human Physiology*, 2012, pp. 701–710.

[3] S. A. McQuilken, “The mouth, stomach and intestines,” *Anaesth. Intensive Care Med.*, vol. 22, no. 5, pp. 330–335, 2021, doi: 10.1016/j.mpaic.2021.04.001.

[4] M. B. Piazuelo, M. Epplein, and P. Correa, “Gastric cancer : An infectious disease,” *HHS Author Manuscripts*, vol. 24, no. 4, pp. 853–869, 2011, doi: 10.1016/j.idc.2010.07.010.

[5] A. R. Yusefi, K. B. Lankarani, P. Bastani, M. Radinmanesh, and Z. Kavosi, “Risk Factors for Gastric Cancer: A Systematic Review,” *Asian Pacific J. Cancer Prev.*, vol. 19, no. 3, pp. 591–603, 2018, doi: 10.22034/APJCP.2018.19.3.591.

[6] H. Zali, M. Rezaei-tavirani, and M. Azodi, “Gastric cancer : prevention , risk factors and treatment,” vol. 4, no. April, pp. 175–185, 2011.

[7] P. Rawla and A. Barsouk, “Epidemiology of gastric cancer : global trends , risk factors and prevention,” *Gastroenterol. Revis.*, vol. 14, no. 1, pp. 26–38, 2019, doi: doi.org/10.5114/pg.2018.80001.

[8] S. J. Kim and C. Woong, “Common Locations of Gastric Cancer: Review of Research from the Endoscopic Submucosal Dissection Era,” *J. Korean Med. Sci.*, vol. 34, no. 35, 2019, doi: 10.3346/jkms.2019.34.e231.

[9] S. Nagini, “Carcinoma of the stomach: A review of epidemiology, pathogenesis, molecular genetics and chemoprevention,” *World J. Gastrointest. Oncol.*, vol. 4, no. 7, pp. 156–169, 2012, doi: 10.4251/wjgo.v4.i7.156.

[10] D. M. Richman *et al.*, “Beyond gastric adenocarcinoma: Multimodality assessment of common and uncommon gastric neoplasms,” *HHS Author Manuscripts*, vol. 176, no. 5, pp. 139–148, 2017, doi: 10.1007/s00261-016-0901-x.Beyond.

[11] E. Morgan *et al.*, “The current and future incidence and mortality of gastric cancer in 185 countries, 2020–40: A population-based modelling study,” *eClinicalMedicine*, vol. 47, pp. 1–10, 2022, doi: 10.1016/j.eclinm.2022.101404.

[12] Y. Song, X. Liu, W. Cheng, H. Li, and D. Zhang, “The global, regional and national burden of stomach cancer and its attributable risk factors from 1990 to 2019,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-15839-7.

[13] F. Carneiro, *Gastric Cancer*. Elsevier Inc., 2014.

[14] S. Morais, A. Ferro, A. Bastos, and C. Castro, “Trends in gastric cancer mortality and in the prevalence of Helicobacter pylori infection in Portugal,” *Eur. J. Cancer Prev.*, vol. 00, no. 00, pp. 1–7, 2015, doi: 10.1097/CEJ.0000000000000183.

[15] K. D. Crew and A. I. Neugut, “Epidemiology of gastric cancer,” *World J. Gastroenterol.*, vol. 12, no. 3, pp. 354–362, 2006, doi: 10.3748/wjg.v12.i3.354.

[16] J. C. Ford and J. A. Ford, “Multimorbidity: will it stand the test of time?,” *Age Ageing*, vol. 47, no. 1, pp. 6–8, 2018, doi: 10.1093/ageing/afx159.

[17] A. Calderon-Larranaga *et al.*, “Rapidly developing multimorbidity and disability in older adults : does social background matter ?,” *J. Intern. Med.*, vol. 283, pp. 489–499, 2018, doi: 10.1111/joim.12739.

[18] A. of M. Sciences, “Multimorbidity: a priority for global health research,” 2018.

[19] K. Nicholson and M. Fortin, “The Measurement of Multimorbidity,” *Heal. Psychol.*, vol. 38, no. 9, 2019, doi: 10.1037/hea0000739.

[20] C. Harrison *et al.*, “Comorbidity versus multimorbidity: Why it matters,” *J. Multimorbidity Comorbidity*, vol. 11, p. 263355652199399, 2021, doi: 10.1177/2633556521993993.

[21] S. Kadambi, M. Abdallah, and K. P. Loh, “Multimorbidity, Function and Cognition in Aging,” *Clin. Geriatr. Med.*, vol. 36, no. 4, pp. 569–584, 2020, doi: 10.1016/j.cger.2020.06.002.

[22] A. Prados-Torres *et al.*, “Multimorbidity Patterns in Primary Care : Interactions among Chronic Diseases Using Factor Analysis,” *PLoS One*, vol. 7, no. 2, 2012, doi: 10.1371/journal.pone.0032190.

[23] M. C. Johnston, M. Crilly, C. Black, G. J. Prescott, and S. W. Mercer, “Defining and measuring multimorbidity: a systematic review of systematic reviews,” *Eur. J. Public Health*, vol. 29, no. 1, pp. 182–189, 2019, doi: doi.org/10.1093/eurpub/cky098.

[24] I. S.-S. Ho, Azcoaga-Lorenzo, A. Amaya Akbari, C. Black, J. Davies, and P. Hodgins, “Examining variation in the measurement of multimorbidity in research: a systematic review of 566 studies,” *Lancet*, vol. 6, no. 8, pp. 587–597, 2021, doi: 10.1016/S2468-2667(21)00107-9.

[25] A. Prados-Torres, A. Calderón-Larrañaga, J. Hancco-Saavedra, B. Poblador-Plou, and M. Van Den Akker, "Multimorbidity patterns: A systematic review," *J. Clin. Epidemiol.*, vol. 67, no. 3, pp. 254–266, 2014, doi: 10.1016/j.jclinepi.2013.09.021.

[26] A. Elixhauser, C. Steiner, and D. R. Harris, "Comorbidity Measures for Use with Administrative Data Author (s): Anne Elixhauser , Claudia Steiner , D . Robert Harris and Rosanna M . Coffey Published by : Lippincott Williams & Wilkins Stable URL : https://www.jstor.org/stable/3766985 Comorbidity Mea," vol. 36, no. 1, pp. 8–27, 1998.

[27] H. Quan, V. Sundararajan, P. Halfon, and A. Fong, "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 Administrative Data," *Med. Care*, vol. 43, no. 11, 2005.

[28] W. H. Organization, "International Statistical Classification of Diseases and Related Health Problems (ICD)," 2022. [Online]. Available: https://www.who.int/standards/classifications/classification-of-diseases.

[29] E. S. Lee *et al.*, "Systematic review on the instruments used for measuring the association of the level of multimorbidity and clinically important outcomes," *BMJ Open*, vol. 11, pp. 1–21, 2021, doi: 10.1136/bmjopen-2020-041219.

[30] A. L. Huntley, R. Johnson, S. Purdy, J. M. Valderas, and C. Salisbury, "Measures of Multimorbidity and Morbidity Burden for Use in Primary Care and Community Settings: A Systematic Review and Guide," *Ann. Fam. Med.*, vol. 10, no. 2, pp. 134–141, 2012, doi: 10.1370/afm.1363.

[31] H. Nguyen, G. Manolova, C. Daskalopoulou, S. Vitoratou, M. Prince, and A. M. Prina, "Prevalence of multimorbidity in community settings: A systematic review and meta-analysis of observational studies," *J. Comorbidity*, vol. 9, pp. 1–15, 2019, doi: 10.1177/2235042x19870934.

[32] C. P. Diederichs, J. Wellmann, D. B. Bartels, U. Ellert, W. Hoffmann, and K. Berger, "How to weight chronic diseases in multimorbidity indices ? Development of a new method on the basis of individual data from five population-based studies," *J. Clin. Epidemiol.*, vol. 65, pp. 679–685, 2012, doi: 10.1016/j.jclinepi.2011.11.006.

[33] G. Cezard, C. T. McHale, F. Sullivan, J. K. F. Bowles, and K. Keenan, "Studying trajectories of multimorbidity : a systematic scoping review of longitudinal approaches and evidence," *BMJ Open*, vol. 11, no. 11, pp. 1–19, 2021, doi: 10.1136/bmjopen-2020-048485.

[34] J. Agterberg, F. Zhong, R. Crabb, and M. Rosenberg, "Cluster analysis application to identify groups of individuals with high health expenditures," *Heal. Serv. Outcomes Res. Methodol.*, vol. 20, pp. 140–182, 2020, doi: Agterberg, Joshua; Zhong, Fanghao; Crabb, Richard; Rosenberg, Marjorie (2020). Cluster analysis application to identify groups of individuals with high health expenditures. Health Services and Outcomes Research Methodology, 20(2-3), 140–182. doi:10.1007/s10742-020-00214-8.

[35] N. Sharma, R. Schwendimann, O. Endrich, D. Ausserhofer, and M. Simon, "Comparing Charlson and Elixhauser comorbidity indices with different weightings to predict in-hospital mortality: an analysis of national inpatient data," *BMC Health Serv. Res.*, vol. 21, no. 13, 2021, doi: 10.1186/s12913-020-05999-5.

[36] C. Diederichs, K. Berger, and D. B. Bartels, "The Measurement of Multiple Chronic Diseases—A Systematic Review on Existing Multimorbidity Indices," *Journals Gerontol. Ser. A*, vol. 66A, no. 3, pp. 301–311, 2011, doi: 10.1093/gerona/glq208.

[37] H. Van Den Bussche *et al.*, "Which chronic diseases and disease combinations are specific to multimorbidity in the elderly? Results of a claims data based cross-sectional study in Germany," *BMC Public Health*, vol. 11, 2011, doi: 10.1186/1471-2458-11-101.

[38] M. Kailasam, W. Guo, Y. M. Hsann, and K. S. Yang, "Prevalence of care fragmentation among outpatients attending specialist clinics in a regional hospital in Singapore: a cross-sectional study," *BMJ Journals*, vol. 9, no. 3, 2019, doi: 10.1136/bmjopen-2018-022965.

[39] I. S.-S. Ho *et al.*, "Variation in the estimated prevalence of multimorbidity: systematic review and meta-analysis of 193 international studies," *BMJ Open*, vol. 12, 2022, doi: 10.1136/ bmjopen-2021-057017.

[40] S. Dekhtyar *et al.*, "Original Contribution Association Between Speed of Multimorbidity Accumulation in Old Age and Life Experiences : A Cohort Study," *Am. J. Epidemiol.*, vol. 188, no. 15, pp. 1627–1636, 2019, doi: 10.1093/aje/kwz101.

[41] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *Lancet*, vol. 380, no. 9836, pp. 37–43, 2012, doi: 10.1016/S0140-6736(12)60240-2.

[42] A. Basto-Abreu *et al.*, "Multimorbidity matters in low and middle-income countries," *J. Multimorbidity Comorbidity*, 2022, doi: 10.1177/2633556522110607.

[43] T. J. Bollyky, T. Templin, M. Cohen, and J. L. Dieleman, "Lower-Income Countries That Face The Most Rapid Shift In Noncommunicable Disease Burden Are Also The Least Prepared," *Health Aff.*, vol. 36, no. 11, 2017, doi: 10.1377/hlthaff.2017.0708.

[44] P. B. Tran, J. Kazibwe, G. F. Nikolaidis, I. Linnosmaa, M. Rijken, and J. van Olmen, "Costs of multimorbidity: a systematic review and meta-analyses," *BMC Med. Vol.*, vol. 20, 2022, doi: 10.1186/s12916-022-02427-9.

[45] D. L. B. Souza *et al.*, “Trends of multimorbidity in 15 European countries: a population-based study in community-dwelling adults aged 50 and over,” *BMC Public Health*, vol. 76, 2021, doi: 10.1186/s12889-020-10084-x.

[46] G. Quinaz Romana, I. Kislaya, M. R. Salvador, S. Cunha Gonçalves, B. Nunes, and C. Dias, “Multimorbidity in Portugal: Results from the first national health examination survey,” *Acta Med. Port.*, vol. 32, no. 1, pp. 30–37, 2019, doi: 10.20344/amp.11227.

[47] F. Prazeres and L. Santiago, “Prevalence of multimorbidity in the adult population attending primary care in Portugal : a cross-sectional study,” *BMJ Journals*, vol. 5, 2015, doi: 10.1136/bmjopen-2015-009287.

[48] A. Hassaine, G. Salimi-Khorshidi, D. Canoy, and K. Rahimi, “Untangling the complexity of multimorbidity with machine learning,” *Mech. Ageing Dev.*, vol. 190, 2020, doi: 10.1016/j.mad.2020.111325.

[49] A. Bisquera *et al.*, “The Lancet Regional Health - Europe Identifying longitudinal clusters of multimorbidity in an urban setting : A population-based cross-sectional study,” *Lancet Reg. Heal. - Eur.*, vol. 3, 2021, doi: 10.1016/j.lanepe.2021.100047.

[50] K. Nicholson, M. Bauer, A. L. Terry, M. Fortin, T. Williamson, and A. Thind, “The multimorbidity cluster analysis tool: identifying combinations and permutations of multiple chronic diseases using a record-level computational analysis,” *BMJ Journals*, vol. 24, 2017, doi: 10.14236/jhi.v24i4.962.

[51] A. R. Quiñones, S. Markwardt, and A. Botoseneanu, “Multimorbidity Combinations and Disability in Older Adults,” *Journals Gerontol.*, vol. 71, no. 6, pp. 823–830, 2016, doi: 10.1093/gerona/glw035.

[52] A. Marengoni, E. Von Strauss, D. Rizzuto, B. Winblad, and L. Fratiglioni, “The impact of chronic multimorbidity and disability on functional decline and survival in elderly persons. A community-based, longitudinal study,” *J. Intern. Med.*, vol. 265, no. 2, pp. 288–295, 2009, doi: 10.1111/j.1365-2796.2008.02017.x.

[53] Q. D. Nguyen, C. Wu, M. C. Odden, and D. H. Kim, “Multimorbidity Patterns, Frailty, and Survival in Community-Dwelling Older Adults,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 78, no. 8, pp. 1265–1270, 2019, doi: 10.1093/gerona/gly205.

[54] S. L. Marrero, D. E. Bloom, and E. Y. Adashi, “Noncommunicable Diseases A Global Health Crisis in a New World Order,” *JAMA Netw. Open*, vol. 307, no. 19, pp. 2037–2038, 2012, doi: 10.1001/jama.2012.3546.

[55] B. PereiraNunes, T. R. Flores, L. A. Facchini, G. Iven Mielke, and E. Thumé, “Multimorbidity and mortality in older adults: A systematic review and meta-analysis,” *Arch. Gerontol. Geriatr.*, vol. 67, pp. 130–138, 2016, doi: doi.org/10.1016/j.archger.2016.07.008.

[56] A. Dhere, “Managing complex long-term conditions and multimorbidity,” *Clin. Med. (Northfield. Il).*, vol. 16, no. 6, 2016, doi: 10.7861/clinmedicine.16-6-545.

[57] Y. H. Chen, M. Karimi, M. P. M. H. R. Mölken, and J. F. Orueta, “The disease burden of multimorbidity and its interaction with educational level,” *PLoS One*, vol. 15, no. 12, 2020, doi: 10.1371/journal.pone.0243275.

[58] K. Moffat and S. W. Mercer, “Challenges of managing people with multimorbidity in today’s healthcare systems,” *BMC Prim. Care*, vol. 129, 2015, doi: 10.1186/s12875-015-0344-4.

[59] H. Fowler *et al.*, “Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers,” *BMC Cancer*, vol. 20, 2020, doi: 10.1186/s12885-019-6472-9.

[60] T. Ahmad, D. Gopal, Z. M. D. Ullah, and S. Taylor, “Multimorbidity in patients living with and beyond cancer: protocol for a scoping review,” *BMJ Journals*, vol. 12, no. 5, 2022, doi: 10.1136/bmjopen-2021-057148.

[61] M. Rijken *et al.*, “How to improve care for people with multimorbidity in Europe ?,” *ICARE4EU project*, 2016. .

[62] K. Togo and N. Yonemoto, “Real world data and data science in medical research: present and future,” 2022.

[63] J. Yang *et al.*, “Brief introduction of medical database and data mining technology in big data era,” *J. Evid. Based. Med.*, vol. 13, no. 1, pp. 57–69, 2020, doi: 10.1111/jebm.12373.

[64] W.-T. Wu, Y.-J. Li, L. L. Ao-Zi Feng, T. Huang, A.-D. Xu, and J. Lyu, “Data mining in clinical big data: the frequently used databases, steps, and methodological models,” *Mil. Med. Res.*, vol. 8, no. 44, 2021, doi: 10.1186/s40779-021-00338-z.

[65] D. Talia, P. Trunfio, and F. Marozzo, “Introduction to Data Mining,” in *Data Analysis in the Cloud*, 2016, pp. 1–25.

[66] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, “Evaluation of cluster validation metrics,” *Comput. Learn. Approaches to Data Anal. Biomed. Appl.*, pp. 189–208, 2020, doi: 10.1016/b978-0-12-814482-4.00007-3.

[67] J. Bible, S. Datta, and S. Datta, *Chapter 4. Cluster Analysis: Finding Groups in Data*, no. 2008. Elsevier Inc., 2013.

[68] A. Roso-llorach *et al.*, “Comparative analysis of methods for identifying multimorbidity patterns : a study of ‘ real-world ’ data,” *BMJ Open*, vol. 8, pp. 1–12, 2018, doi: 10.1136/bmjopen-2017-018986.

[69] M. Tavakol and A. Wetzel, “Factor Analysis: a means for theory and instrument development in support of construct

validity," *Int. J. Med. Educ.*, vol. 11, pp. 245–247, 2020, doi: 10.5116/ijme.5f96.0f4a.

[70] D. Hevey, "Network analysis: a brief overview and tutorial," *Heal. Psychol. Behav. Med.*, vol. 6, no. 1, 2018, doi: 10.1080/21642850.2018.1521283.

[71] Y. Chen and R. Xu, "Network Analysis of Human Disease Comorbidity Patterns Based on Large-Scale Data Mining," in *Lecture Notes in Computer Science*, 2014, pp. 243–254.

[72] A. Amelio and A. Tagarelli, "Data Mining : Clustering," in *Encyclopedia of Bioinformatics and Computational Biology*, 2018.

[73] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017, doi: 10.1016/j.neucom.2017.06.053.

[74] I. Kononenko and M. Kukar, "Cluster Analysis," in *Machine Learning and Data Mining*, 2007, pp. 321–358.

[75] X. Jin and J. Han, "K-Medoids Clustering," *Encyclopedia of Machine Learning*. pp. 564–565, 2011, doi: 10.1007/978-0-387-30164-8_426.

[76] D. Coomans, C. Smyth, I. Lee, T. Hancock, and J. Yang, "Unsupervised Data Mining: Introduction," *Compr. Chemom.*, vol. 2, pp. 559–576, 2009, doi: 10.1016/B978-044452701-1.00063-6.

[77] S. Mahmood, M. Shahbaz, and A. Guergachi, "Negative and Positive Association Rules Mining from Text Using Frequent and Infrequent Itemsets," vol. 2014, 2014, doi: 10.1155/2014/973750.

[78] E. E. Services, "Advanced Analytical Theory and Methods," in *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, 2015, p. Association Rules.

[79] K. W. Siah, C. H. Wong, J. Gupta, and A. W. Lo, "Multimorbidity and mortality: A data science perspective," *J. Multimorbidity Comorbidity*, vol. 12, 2022, doi: 10.1177/26335565221105431.

[80] F. P. Held *et al.*, "Association Rules Analysis of Comorbidity and Multimorbidity: The Concord Health and Aging in Men Project," *Journals Gerontol. Ser. A*, vol. 71, no. 5, pp. 625–631, 2016, doi: 10.1093/gerona/glv181.

[81] D. L. B. de Souza *et al.*, "Multimorbidity and its associated factors among adults aged 50 and over: A cross-sectional study in 17 European countries," *PLoS One*, vol. 16, no. 2, 2021, doi: 10.1371/journal.pone.0246623.

[82] L. Robertson, R. Vieira, J. Butler, M. Johnston, S. Sawhney, and C. Black, "Identifying multimorbidity clusters in an unselected population of hospitalised patients International Classification of Diseases," *Sci. Rep.*, vol. 12, pp. 1–10, 2022, doi: 10.1038/s41598-022-08690-3.

[83] Y. Lee, H. Kim, and H. Jeong, "Patterns of Multimorbidity in Adults : An Association Rules Analysis Using the Korea Health Panel," *Int. J. Environ. Res. Public Health*, vol. 17, no. 8, 2020, doi: 10.3390/ijerph17082618.

[84] J. Marley, B. I. Nicholl, S. Macdonald, F. S. Mair, and B. D. Jani, "Associations between long-term conditions and upper gastrointestinal cancer incidence: A prospective population-based cohort of UK Biobank participants," *SAGE JOURNALS*, vol. 11, pp. 1–13, 2021, doi: 10.1177/26335565211056136.

[85] H. Wickham, R. François, L. Henry, and K. Müller, "dplyr: A Grammar of Data Manipulation.," 2022. [Online]. Available: https://cran.r-project.org/web/packages/dplyr/index.html.

[86] D. D. Sjoberg *et al.*, "gtsummary: Presentation-Ready Data Summary and Analytic Result Tables," 2022. [Online]. Available: https://cran.r-project.org/web/packages/gtsummary/index.html.

[87] M. Maechler *et al.*, "cluster: 'Finding Groups in Data': Cluster Analysis Extended Rousseeuw et al.," 2022. [Online]. Available: https://cran.r-project.org/web/packages/cluster/index.html.

[88] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP J. Wirel. Commun. Netw.*, 2021, doi: 10.1186/s13638-021-01910-w.

[89] K. Bolar, "STAT: Interactive Document for Working with Basic Statistical Analysis," 2019. [Online]. Available: https://cran.r-project.org/web/packages/STAT/index.html.

[90] C. Hennig, "fpc: Flexible Procedures for Clustering," 2020. [Online]. Available: https://cran.r-project.org/web/packages/fpc/index.html.

[91] G. Brock, V. Pihur, S. Datta, and S. Datta, "clValid: Validation of Clustering Results," 2021. [Online]. Available: https://cran.r-project.org/web/packages/clValid/index.html.

[92] M. Hahsler *et al.*, "arules: Mining Association Rules and Frequent Itemsets," 2022. [Online]. Available: https://cran.r-project.org/web/packages/arules/index.html.

[93] P. Osisikankwu and L. Ume, "Discovery Of Associated Items Using Apriori Algorithm For Chain Stores," no. June, pp. 0–10, 2020.

[94] S. Epskamp and E. I. Fried, “bootnet: Bootstrap Methods for Various Network Estimation Routines,” 2021. [Online]. Available: https://cran.r-project.org/web/packages/bootnet/index.html.

[95] D. E. da C. Leme, E. V. da C. Alves, V. do C. O. Lemos, and A. Fattori, “Network Analysis: a Multivariate Statistical Approach for Health Science Research,” *Geriatr. Gerontol. Aging*, vol. 14, no. 1, pp. 43–51, 2020, doi: 10.5327/z2447-212320201900073.

[96] B. M. Bolker *et al.*, “Generalized linear mixed models : a practical guide for ecology and evolution,” no. Table 1, pp. 127–135, 2008, doi: 10.1016/j.tree.2008.10.008.

[97] Eurostat, “PORTUGAL-NUTSlevel2,” 2020. [Online]. Available: https://ec.europa.eu/eurostat/documents/345175/7451602/2021-NUTS-2-map-PT.pdf.

[98] I. Bakbergenuly and E. Kulinskaya, “Meta-analysis of binary outcomes via generalized linear mixed models : a simulation study,” *BMC Med. Res. Methodol.*, vol. 70, pp. 1–18, 2018, doi: 10.1186/s12874-018-0531-9.

[99] S. Nakagawa, P. C. D. Johnson, and H. Schielzeth, “The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded,” *J. R. Soc. Interface*, vol. 14, no. 134, 2017, doi: 10.1098/rsif.2017.0213.

[100] L. Lou *et al.*, “Sex difference in incidence of gastric cancer: An international comparative study based on the Global Burden of Disease Study 2017,” *BMJ Open*, vol. 10, no. 1, pp. 1–7, 2020, doi: 10.1136/bmjopen-2019-033323.

[101] P. Broeiro-gonçalves, P. Nogueira, and P. Aguiar, “Multimorbidity and Disease Severity Measured by the Charlson Index in Portuguese Hospitalised Patients During the Year 2015 : A Cross-Sectional Study,” *Acta Med. Port.*, vol. 32, no. 1, pp. 38–46, 2019, doi: 10.20344/amp.9728.

[102] A. Legler, E. H. Bradley, and M. D. A. Carlson, “The effect of comorbidity burden on health care utilization for patients with cancer using hospice,” *J. Palliat. Med.*, vol. 14, no. 6, pp. 751–6, 2011, doi: 10.1089/jpm.2010.0504.

[103] W. C. Meijers and R. A. De Boer, “Common risk factors for heart failure and cancer,” *Eur. Soc. Cardiol.*, vol. 115, pp. 844–853, 2019, doi: 10.1093/cvr/cvz035.

[104] D. E. V. Olivares, F. R. V. Chambi, E. M. M. Chañi, W. J. Craig, S. O. S. Pacheco, and F. J. Pacheco, “Risk Factors for Chronic Diseases and Multimorbidity in a Primary Care Context of Central Argentina: A Web-Based Interactive and Cross-Sectional Study,” *Environ. Res. Public Heal.*, vol. 14, no. 3, p. 251, 2017, doi: 10.3390/ijerph14030251.

[105] C. Tseng and F. Tseng, “Diabetes and gastric cancer : the potential links,” *World J. Gastroenterol.*, vol. 20, no. 7, pp. 1701–1711, 2014, doi: 10.3748/wjg.v20.i7.1701.

[106] S. M. Dunlay, S. A. Weston, S. J. Jacobsen, and V. L. Roger, “Risk Factors for Heart Failure: A Population-Based Case-Control Study,” *NIH Public Access*, vol. 122, no. 11, pp. 1023–1028, 2010, doi: 10.1016/j.amjmed.2009.04.022.Risk.

[107] S. Osman, C. Ziegler, R. Gibson, R. Mahmood, and J. Moraros, “The Association between Risk Factors and Chronic Obstructive Pulmonary Disease in Canada : A Cross - sectional Study Using the 2014 Canadian Community Health Survey,” *Int. J. Prev. Med.*, vol. 8, no. 86, 2017, doi: 10.4103/ijpvm.IJPVM.

[108] S. A. Mahmoodi, K. Mirzaie, and S. M. Mahmoudi, “A new algorithm to extract hidden rules of gastric cancer data based on ontology,” *Springerplus*, vol. 5, no. 312, 2016, doi: 10.1186/s40064-016-1943-9.

[109] C. Castro, B. Peleteiro, M. J. Bento, and N. Lunet, “Trends in gastric and esophageal cancer incidence in northern Portugal (1994-2009) by subsite and histology, and predictions for 2015,” *Br. J. Cancer*, vol. 103, no. 2, pp. 155–163, 2017, doi: 10.5301/tj.5000542.

[110] C. Castro, B. Peleteiro, M. J. Bento, and N. Lunet, “Prediagnosis lifestyle exposures and survival of gastric cancer patients : a cohort study from Portugal,” *Br. J. Cancer*, vol. 107, no. 2, pp. 537–543, 2012, doi: 10.1038/bjc.2012.258.

[111] J. Souza *et al.*, “Measuring Variability in Acute Myocardial Infarction Coding Using a Statistical Process Control and Probabilistic Temporal Data Quality Control Approaches,” in *Trends and Applications in Information Systems and Technologies*, 2021, pp. 193–202, doi: 10.1007/978-3-030-72651-5_19.