Identifying Subgroups in Heart Failure Patients with Multimorbidity by Clustering Analysis

Catarina Patrício Mendes Martins, catarina.p.martins@tecnico.ulisboa.pt[1]

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

***Abstract*—This work presents a workflow for the identification and characterisation of clinically significant Heart Failure (HF) patient subgroups with multimorbidity using data from Electronic Health Records. Multimorbidity, defined as the co-occurrence of two or more chronic conditions, represents a heavy burden for healthcare systems. However, there is still a lack of knowledge about patients with multimorbidity, the most common disease interactions, risk factors and treatment response. This is particularly relevant in complex and heterogeneous conditions such as HF. Using clinical data from Hospital da Luz Lisboa, we conducted a clustering analysis of HF patients and characterised the clusters in terms of demographics, comorbidities, laboratory values, medical prescriptions and medical appointments. This was followed by a survival analysis to identify associations between clusters and outcomes such as hospitalisation and emergency admission. The workflow identified four distinct clusters, with significant differences in the clinical variables analysed. The survival analysis showed differential associations for hospitalisation and emergency admission risks between clusters, and the risk associations were in agreement with the cluster profiles. The results evidence a high degree of disease heterogeneity within HF patients and how an improved characterisation of patient subgroups can be relevant for clinical risk stratification.**

***Index Terms*—Multimorbidity, Heart Failure, Electronic Health Records, Phenotyping, Clustering**

I. Introduction

As life expectancy increases, the percentage of people suffering from more than one chronic condition is increasing dramatically. This co-occurrence of two or more chronic conditions, defined as multimorbidity, is estimated to affect around 50 million people in the European Union, making it one of the most challenging problems faced by the health sector at the current time [1].

Patients with multimorbidity pose a challenge in many ways. Patients with more than one chronic condition are associated with substantially higher health care utilisation and costs [2]. Additionally, multimorbidity is linked to more complex medical care, poorer health outcomes and quality of life, and faster disease progression [1]. Despite all negative impacts, current healthcare systems are not prepared to cope with patients with multimorbidity [3]. It is necessary to change healthcare from a disease-oriented approach to a patient-centred multidisciplinary care, which takes into account the different diseases and the possible interactions among them.

Thus, there is a need to study and characterise populations of patients with multimorbidity. The identification of the patients' conditions and specific characteristics, a process denominated as phenotyping, can bring several advantages. The characterisation and phenotyping of patient cohorts can provide a better understanding of the most common associations and interactions between diseases, providing helpful insights for predicting treatment response, and aiding in the design of drug development strategies and clinical trials [4, 5].

The Electronic Health Record (EHR) is the standardised tool for capturing patients' medical history and constitutes a key source of information for patient phenotyping. EHRs contains structured data, such as laboratory results and diagnoses, along with unstructured data, such as radiology reports, discharge summaries, and other clinical narratives. Information from structured data can be easily used in phenotyping algorithms, whereas information from unstructured data needs to be extracted and analysed, which poses a challenge by itself [6].

There are several approaches to phenotyping. In recent years, research has been shifting to machine learning methodologies [6]. Unsupervised learning algorithms, such as clustering, have the advantage of not needing an *a priori* classification, which means the clusters obtained are based only on the patterns found in the data.

A relevant application for characterising patient cohorts is related to the identification and phenotyping of patient subgroups within a specific conditions. Certain conditions, such as Heart Failure (HF), Chronic Obstructive Pulmonary Disease (COPD) and Parkinson's have distinct subtypes [7, 8, 9]. Proper characterisation of patients within each subtype has an impact on therapeutic response and patient outcomes. Several studies have shown that such conditions can benefit greatly from identifying specific patient cohorts since this information can aid in clinical decision making and assess risks for each cohort [10, 11].

HF is estimated to affect 64.3 million people worldwide, and in developed countries its prevalence is generally estimated at 1% to 2% of the general adult population [12]. Despite some improvements over the last years, HF prognosis remains poor and patients' quality of life remains low [13]. The high complexity of HF syndrome and high heterogeneity of patients makes disease management extremely challenging, and can lead to ineffective treatment in some cases. There is a need for improved characterisation of HF patients, that can help design clinical trials and contribute to defining therapeutic strategies appropriate for the patients' characteristics.

II. Concepts and Related Work

In the past years, EHRs have been adopted by a vast number of countries which has generated a large amount of clinical data. One of the most promising applications of EHRs is to identify groups of patients with certain characteristics, a process named electronic phenotyping.

The characterisation of patient cohorts and study of the corresponding phenotypes can provide a better understanding of the most common associations and interactions between diseases, providing helpful insights for predicting clinical outcomes. Shivade et al. pointed out that better characterised cohorts can also have an important impact in drug development strategies and clinical trials, with additional information regarding the target population [6].

Over the years different approaches have been developed to identify patient' cohorts using data from the EHRs. Complex approaches have been developed, making use of emerging technologies, such as Machine Learning (ML) and Natural Language Processing (NLP), and integrating clinical information from unstructured data.

Clustering is one of the ML algorithms used in EHRs Phenotyping. In the clinical field, clustering has been used to identify patient groups with similar characteristics, in a general population or within a single disease.

EHRs can contain very heterogeneous data, a mixture of categorical and continuous data. Foss et al. presents a review of some of the methods identified in the literature to deal with mixed-type data [14]. One of the strategies mentioned was using hybrid distance approaches, that is, using specific distance functions prepared for mixed-type data before applying clustering, such as Gower's distance [15], which was used by Singh et al. for clustering patients with multimorbidity [16]. Other approaches include performing data transformation approaches, as discretisation or dimensionality reduction, such as Factor Analysis of Mixed Data (FAMD) [17].

Clustering has been widely applied to populations with multimorbidity, in an attempt to find disease groups that co-occur more frequently.Vetrano et al. studied a population of patients with multimorbidity during 12 years, identifying multimorbidity clusters, detecting clinical trajectories and tracing the clusters evolution [18]. Using Fuzzy C-Means, the study found five clinically meaningful clusters and one unspecified cluster. It was observed that throughout the years the patients in the unspecified cluster would move to one of the other clusters as other diseases developed. They concluded that multimorbidity clusters can help to identify patient groups with similar prognosis and treatment needs and thus assist healthcare professionals in designing appropriate treatment plans and targeting preventive strategies.

In another study, patients with multimorbidity were divided by sex and clustered using k-means on EHRs data [19]. The clustering resulted in six multimorbidity patterns for each gender, with the most prevalent pattern including coincident diseases for both male and female patients. The phenotypes defined by the clustering analysis were consistent with clinical practice and presented similarities to previous studies.

As for discovering phenotypes within a single disease, several studies have used clustering methods to identify clinically relevant patient subgroups in complex syndromes such as HF, COPD, Dementia and Parkinson's Disease, leading to important insights regarding disease pathophysiology [20].

HF is a complex clinical syndrome characterised by the inability of the heart to maintain a cardiac output that suffices the body demands of oxygen and blood. As described by The American College of Cardiology Foundation/American Heart Association, the diagnosis of HF is not straightforward as there is no single diagnostic test, it is predominantly a clinical diagnosis based on patient history and physical examination [21]. The recommended diagnostic tests consist of a chest X-ray, echocardiography, and plasma levels of BNP or NT-proBNP. BNP is a hormone released by the ventricles whenever the heart undergoes stress, whether chronic or acute, in an attempt to compensate the vasoconstrictor systems that are activated in these situations. It has great prognostic value in the context of HF and is a powerful rule out-test. Echocardiogram, for its availability, is considered the key tool for the assessment of cardiac dysfunction.

In the new Universal definition of HF [22], the condition is classified in 4 phenotypes based on the Ejection Fraction (EF): *i)* Heart Failure with Reduced Ejection Fraction (HFrEF), when Left Ventricle Ejection Fraction (LVEF) < 40%, *ii)* Heart Failure with Mildly Reduced Ejection Fraction (HFmrEF), when LVEF is 41-49%, *iii)* Heart Failure with Preserved Ejection Fraction (HFpEF), when LVEF >50% and *iv)* HF with recovered EF when LVEF > 40%. EF is a measurement, usually obtained using an echocardiogram, of the percentage of blood that the heart pumps out with each contraction. A healthy individual usually has a preserved EF, normally above 70%.

There have been several studies focusing on HF phenotyping. Ahmad et al. applied Ward's hierarchical clustering to 2,331 HF patients and identified four clusters whose patients characteristics varied greatly in measures of demographics, symptoms, comorbidities, HF aetiology, quality of life, among others [20]. The study also found differential associations for hospitalisation and mortality risks between clusters. Gulea et al. used Latent Class Analysis to cluster HF patients based only on their comorbidities and obtained five clusters that exhibited differences in the risks of hospital admission, mortality, and healthcare resource utilisation [23].

To compare patient subgroups in terms of risk association with outcomes such as hospitalisation and mortality, Ahmad et al. and Gulea et al. use methods from Survival Analysis, a branch of statistics that focus on studying the lifespan of a particular population under examination [24]. The goal of Survival Analysis is to estimate the expected duration for an individual to experience an event of interest. Survival Analysis provides methods that allow comparing the risk for an event of interest for different groups. The Kaplan–Meier model estimates the survival curve, the log-rank test compares two groups statistically, and Cox's hazards model allows for the inclusion of additional variables [25].

When clustering data with a high number of features, it becomes hard to visualise the resulting clusters and to understand if they present different characteristics and to what degree do they differ. In this work, we use concepts from network science [26], in particular, graphs, as a visualisation tool, to have a better understanding of the patients' comorbidities prevalence and associations, and to study the interaction between medical appointments.

Networks are a simple representation of how entities connect and interact with each other, helping to reveal patterns

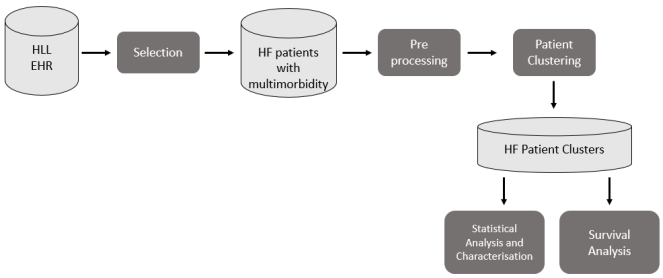


Fig. 1. Overview of the proposed approach for the identification and characterisation of HF patient subgroups using EHRs data from HLL

and providing useful visualisation options. Networks may be represented as graphs, structures composed of nodes (a single point) and edges (connections between points). Edges can be directed (symmetric) or undirected (asymmetric), and can also have an associated weight, that is, a measure of the strength of a link between two nodes. Other important graph concepts are the node degree, which is the number of connections of a node, and the clustering coefficient, which measures the degree to which nodes in a graph tend to cluster together [27].

Hidalgo et al. introduces the concept of Phenotypic Disease Networks (PDNs), a network representation of comorbidities that can be used to study their associations, differences in phenotypes between patients and disease progression. In a PDN, nodes represent diseases and edges are weighted and represent a link between diseases. The weight can be quantified using different measures, such as co-occurrence frequency or Pearson correlation. These networks have the ability to reveal non-obvious relationships between comorbidities that could bring important information to improve patient treatment approaches.

III. Methods

In this work, we developed a workflow for the identification and characterisation of HF patient subgroups. The workflow is represented in Figure 1 and consists of the following steps: (1) preprocessing of the data and exploratory data analysis to characterise the dataset; (2) clustering; (3) statistical analysis for characterisation and visualisation of the obtained clusters, (4) survival analysis to stratify the patient clusters according to the risk for a given outcome.

A. Data

The dataset used to develop the pipeline was generated from an initial population of 54 827 patients, with an observation period between January 2007 and August 2021. From this initial pool, 3745 HF patients with multimorbidity were identified using ICD-9 codes for HF and heart disease, or searching for associated keywords in their medical records (see Table I).

For the clustering analysis, relevant features were selected based on literature review and guidance from the HF specialist at HLL. These features consisted of clinical variables, demographics, physical characteristics, laboratory data and the most common comorbidities associated with HF, amounting to a total of 35 features (see Table II). Lifestyle-related features and gender were provided as text fields, comorbidities as binary variables based on the presence or absence of the disease, and all others as numeric. Besides the features used for clustering, we also extracted the date of HF diagnosis and gathered data on other comorbidities, prescriptions (medications ordered for each patient), and clinical outcomes, namely hospitalisations, emergency admissions and mortality. Data for prescriptions are relative to a period of 9 years and 8 months, from January 2012 until August 2021. For outcomes of hospitalisation and emergency admissions, it was possible to obtain dated records. For the outcome mortality, it was not possible to obtain dated information.

TABLE I
ICD-9 Codes and keywords used to identify HF patients from the HLL database. ICFEp - Insuficiência Cardíaca com Fração de Ejeção Preservada, ICFEr - Insuficiência Cardíaca com Fração de Ejeção Reduzida, IC - Insuficiência Cardíaca, ICC- Insuficiência Cardíaca Congestiva, NYHA - New York Heart Association.

ICD-9 Codes	428, 398.9.1, 402.0.1, 402.9.1, 404.0.1, 404.0.3, 404.1.1, 404.1.3, 404.9.1, 404.9.3, 425.4, 425.5, 425.6, 425.7, 425.8, 425.9
Keywords	Insuficiência cardíaca, Insuficiência cardiaca, Insuficiencia cardiaca, Insuficiencia cardíaca, ICFEp, ICFEr, IC, ICC, NYHA

TABLE II
Feature summary of the Hospital da Luz Lisboa HF dataset.

Phenotypic Domain	Phenotypes
Demographics	**Age**, **Gender**
Physical Characteristics	**Body mass index (BMI)**
Lifestyle	Drug use, Alcohol use, Smoking habits
Laboratory	**Sodium**, Potassium, Bicarbonate, **Urea**, **Creatinine**, GFR, Fasting Glucose, **Hemoglobin**, **Platelet count**, **RDW** ,**NT-proBNP**, Ferritin, Uric Acid, Sedimentation Rate
Comorbidities	**Ischemic Cardiomyopathy**, **Hypertension**, **Diabetes**, **Atrial Fibrillation**, **Cerebrovascular Disease**, **Valvular Disease**, **Chronic Kidney Disease**, **Anaemia**, **Chronic Obstructive Pulmonar Disease**, **Obesity**
Patient complexity	**Number of non-chronic diseases**, **Number of chronic diseases**, **Number of ICD-9 codes**, **Number of consultations**

B. Preprocessing

Prior to the clustering, it was necessary to preprocess the dataset obtained from the EHRs extraction. Categorical features were converted into numeric binary features and features with a prevalence lower than 2% in the cohort were removed. Features with a percentage of missing values higher than 40% were deleted. According to Waljee et al. (2013), the two methods that resulted in the least imputation error and prediction difference when applied to a dataset containing laboratory data were missForest and multivariate imputation by chained equations (MICE) [29]. Missing values were imputed using Python's function *Iterative Imputer*, which is based on the MICE method. The MICE method models the missing values of each feature as a function of other features [30]. To do that, at each step, one of the feature columns is designated output y and the rest of the feature columns are designated as inputs

x. To cover for possible coding errors, the feature Anaemia was defined based on the value of Hemoglobin (Hemoglobin < 12 for woman and Hemoglobin < 13.5 for men [31]) and the feature obesity feature was defined based on the value of the feature BMI (BMI > 30 [32]). Continuous features were normalised to have a mean of 0 and a standard deviation of 1 (using Python's function *StandardScaler*). Categorical binary features were scaled from $\{0, 1\} \rightarrow \{-0.5, 0.5\}$. After preprocessing the total number of features used for clustering was 25. The features are identified in Table II in bold.

In addition to the preprocessing, an Exploratory Data Analysis (EDA) was also performed. EDA uses summary statistics and graphical representations to analyse datasets and get a better comprehension of the data. Histograms and boxplots were used to visualise the distribution of continuous variables and barplots were used to visualise the distribution of categorical variables, such as the distribution of comorbidities. Continuous data were described using median (25th -75th quantile) and categorical data using percentages.

C. Patient Clustering

To apply clustering to the HF dataset from HLL, it was first necessary to determine which clustering algorithm and possible complementary techniques to use, and how to evaluate the clustering. It also was necessary to take into account that the dataset was composed of both numerical and categorical data. Several clustering algorithms were tested to understand which one would be more suitable for the HF data.

Based on the literature regarding clustering mixed-type data and on HF clustering, we tested the following combinations of methods for this dataset:

1) Gower's distance matrix [15] together with Ward's Agglomerative Hierarchical Clustering, [33]
2) dimensionality reduction followed by Ward's Agglomerative Hierarchical Clustering [33];
3) dimensionality reduction followed by K-Means [34]

The dimensionality reduction method chosen was Factor Analysis of Mixed Data (FAMD) [17], a principal component method specific to analyse quantitative and qualitative variables.

Considering that clustering HF patients is an exploratory analysis, there are no ground truth labels and thus only internal validity indices, that depend uniquely on the data being clustered and on the resulting labels, could be used. The used indices were Silhouette Score [35], Calinski-Harabasz [36], and Davies-Bouldin [37].

After computing the three indices for the different clustering algorithms and different values of k, we chose the best method using a majority vote, that is, the algorithm and k that performed best in at least two of the indices. Additionally, as in Ahmad et al., a minimum of $N > 375$ was also defined to promote stability and ensure that there was not any cluster with less than 10% of the total population.

D. Statistical Analysis and Characterisation of Obtained Clusters

After choosing the number of clusters and obtaining the clusters, it was necessary to characterise the patient subgroups, evaluate if they had statistically significant differences, and determine adequate methods to help visualise the clusters. Following the work by Gulea et al., demographic, clinical and laboratory characteristics were compared between groups using Chi-squared tests for categorical variables and Kruskal-Wallis test for continuous variables, computing the respective p-values [23]. We characterised Clusters according to age, gender, and most predominant comorbidities.

To have a better visualisation of the clusters comorbidities prevalence, and associations, we computed a graph representation of each cluster. In this case, each graph's node represents a disease in the cluster and an edge represents a co-occurrence of the two nodes (diseases) connected by that edge. The graphs were created with Python's *NetworkX*package and *Gephi*for visualisation. The settings were adjusted so that node size was proportional to the number of connections to other nodes (node degree) and edge thickness was proportional to disease co-occurrence prevalence (edge weight). Co-occurrences (edges) with a co-occurrence prevalence lower than 2% were discarded to declutter the visualisation. The graph representation of the subgroup comorbidities provides a better understanding of which diseases co-occur more frequently and which diseases have the most connections with other diseases.

The obtained patient subgroups were also discussed with a HF specialist, to understand if they were clinically meaningful, and comparable to previous HF clustering studies.

Prescriptions

We collected information on medication prescriptions, which included prescription dates, the speciality of the health professional, and the commercial and common names of the medication prescribed. Following treatment guidelines from the European Society of Cardiology [38] for HF and advisory from the HF specialist, we chose the most relevant medication groups to be analysed. These groups included both HF specific medications, such as Beta-blockers, ACEi and ARB, and medications for the comorbidities found in the dataset, for example, Metformin for type-2 diabetes or Bronchodilators for COPD. We compared the prevalence of the medication groups in each cluster, having into account the most common comorbidities in the clusters. The distribution of medication groups between clusters was compared with Chi-squared tests, computing the respective p-value.

Outcomes

We examined hospitalisations, emergency admissions to the hospital and mortality. A hospitalisation was considered to be any admission to the hospital requiring an overnight stay. Hospitalisations are often planned, but can also occur after appointments, examinations or emergency admissions. An emergency admission, in this case, represents an unplanned, often urgent admission, which occurs when a patient is admitted at the earliest possible time. For the outcomes of hospitalisation and emergency admission, we computed the incidence per year, the percentage of patients that had an incidence in the first year after HF diagnosis, and the percentage of patients that

had an incidence at least one during the time period analysed. For the outcome mortality, we computed the incidence in each cluster. The values obtained were compared using Chi-squared tests and computing the respective p-values.

Survival Analysis

We conducted a Survival Analysis to analyse the relationship between the clusters and the evolution of the outcomes since diagnosis, and to understand if belonging to a certain disease subgroup would mean the patient is at a higher risk of a certain outcome in the future. Survival Analysis provides a direct comparison between clusters, through the survival curves and Hazards Ratio (HR) [24]. As temporal information regarding the outcomes was only available for outcomes of hospitalisation and emergency admission, the outcome mortality was not included in this analysis. We performed two separate analyses for the outcomes of hospitalisation and emergency.

To perform a Survival Analysis, it is necessary to define the starting point that will be common to all patients. In this study, the starting point, or t_0, was defined as the moment of diagnosis and all intervals for the survival analysis were calculated in relation to that moment. The first step was to compute the time intervals between diagnosis and the occurrence of an outcome for all patients. Patients that did not experience the outcome were labelled as censored. Following this and using Python's package *Lifelines*, Kaplan-Meier curves and Cox proportional regression models were computed. Kaplan-Meier curves were computed for each outcome individually and were stratified per cluster, with differences between groups tested using the log-rank test. For Cox proportional regression, following what was done by Shah et al., three different models were used for each outcome: an unadjusted model that only took into account the clusters, a second model adjusted for age and gender and a third model adjusted for age, gender and the laboratory value NT-proBNP, which is often used as a risk marker for HF [39]. HR from Cox regression models are presented in relation to the lowest risk cluster (determined by the lowest percentage of outcomes).

IV. Results and Discussion

A. Characterisation of the Heart Failure Dataset

HF patients are usually complex patients with advanced age and with a high number of other diseases. The HLL dataset includes 3745 records of HF patients with multimorbidity. The median age is 82 years (inter-quartile range 73-88), the number of female patients is 1979 (52.84%), the number of male patients is 1766 (47.16%), and the median BMI is 24.8, as can be observed in Figure 2. The median number of chronic diseases is 5 (inter-quartile range 3-7) with approximately 40% of the population having between 3 and 5 comorbidities and approximately 30% having 6 to 8 comorbidities.

The most common comorbidity in the dataset is Hypertension (HT) (56.58%), followed by Anaemia (53.75%) and Atrial Fibrillation (AFIB) (33.78%). Other diseases with a high prevalence are Chronic Kidney Disease (CKD) (24.94%) and Ischaemic Cardiomyopathy (ICM) (24.09%) (see Figure 3a).

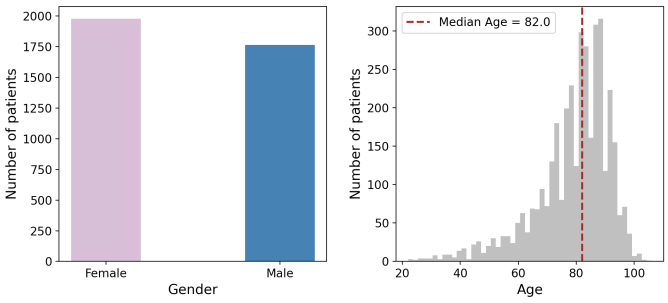
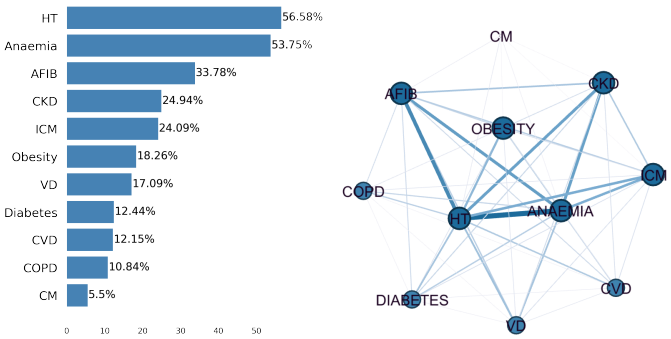


Fig. 2. Gender and age distribution of the dataset of HF patients with multimorbidity



(a) Prevalence of comorbidities.

(b) Graph representation of comorbidities.

Fig. 3. Prevalence and graph representation of comorbidities used for clustering in the HF dataset. In the graph a node represents a disease and its size is proportional to the node degree. An edge represents a co-occurrence of two diseases and its width is proportional to the prevalence of the co-occurrence in the dataset. CM–Ischaemic Cardiomyopathy, HT–Hypertension, AFIB–Atrial Fibrillation, CVD–Cerebrovascular Disease, VD-Valvular Disease, CKD–Chronic Kidney Disease, COPD–Chronic Obstructive Pulmonary Disease

In addition to assessing the prevalence of each comorbidity, a graph representation of the comorbidities was also computed (see Figure 3b). In the representation used, the size of the node is proportional to the number of other nodes it is connected to (node degree), and the width of the edge is proportional to the percentage of co-occurrences of the connected diseases. Using this type of visualisation, it is possible to obtain extra insights into the relationship between comorbidities, as it is possible to observe which diseases tend to co-occur more frequently. In this population, HT, AFIB and ICM are the diseases with higher prevalence and the ones that co-occur more frequently with other diseases. The thickness of the edges makes it possible to verify that HT and Anaemia, and HT and AFIB occur frequently together. CKD and ICM also show a high co-occurrence with HT and Anaemia.

B. Clustering Algorithm and Choice of k

To understand which clustering algorithm would be more suitable for the HF dataset, we used three different approaches, all taking into consideration the mixed-type nature of the data. The results from the different clustering algorithms applied to the HF dataset are summarised in Table III. The combination

TABLE III
VALUES OBTAINED FOR CLUSTERING METRICS SILHOUETTE SCORE. CALINSKI-HARABASZ INDEX AND DAVIES-BOULDIN SCORE FOR THE DIFFERENT CLUSTERING ALGORITHMS, NAMELY, GOWER'S DISTANCE AND HIERARCHICAL CLUSTERING, FAMD AND HIERARCHICAL CLUSTERING, AND FAMD AND K-MEANS. FOR SILHOUETTE SCORE AND CALINSKI-HARABASZ A HIGHER VALUE INDICATES A BETTER PERFORMANCE, FOR DAVIES-BOULDIN A LOWER VALUE IS BEST.

Clustering Algorithm	k	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
Gower Distance + Hierarchical Clustering	3	0.153	918.620	1.859
	4	0.155	910.162	1.751
	5	0.153	810.238	1.797
FAMD + Hierarchical Clustering	3	0.080	221.906	2.267
	4	0.082	225.429	2.325
	5	0.075	227.625	2.332
FAMD + K-Means	3	0.073	217.615	2.374
	4	0.078	201.213	2.420
	5	0.068	191.439	2.537

TABLE IV
VALUES OBTAINED FOR CLUSTERING METRICS SILHOUETTE SCORE. CALINSKI-HARABASZ INDEX AND DAVIES-BOULDIN SCORE FOR HIERARCHICAL CLUSTERING WITH GOWER'S DISTANCE USING K=[2.10]. FOR SILHOUETTE SCORE AND CALINSKI-HARABASZ A HIGHER VALUE INDICATES A BETTER PERFORMANCE, FOR DAVIES-BOULDIN A LOWER VALUE IS BEST.

Clusters	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
2	0.273	1147.505	1.494
3	0.153	918.620	1.859
4	**0.155**	**910.162**	**1.751**
5	0.153	810.238	1.797
6	0.147	746.957	1.739
7	0.141	673.593	1.983
8	0.127	615.439	2.005
9	0.125	571.963	2.041
10	0.130	538.409	1.939
11	0.132	512.288	1.961
12	0.140	491.961	1.914

of Gower's distance matrix and Ward's Hierarchical Agglomerative Clustering is the one that produced clusters with higher scores in all clustering metrics analysed.

The choice of the number of clusters (k) was based on several factors. A minimum of 375 patients was set to promote stability; the metrics Silhouette Score, Calinski-Harabasz, and Davies-Bouldin were analysed; and difference in the resulting cluster characteristics was also evaluated. The metrics score for each value of k was considered by a majority vote, that is, the best value of k is the one that has the highest score in the majority of the metrics considered. Table IV shows the clustering evaluation metrics for clusters with $k = [2, 12]$. The values were the best for $k = 2$. However, the resulting clusters included a cluster with a very high number of patients and a cluster with few patients(under 375). There was also no clear differentiation between patient groups. The second value of k with the best scores was $k = 4$. It was also observed that the resulting clusters had statistically different characteristics and were also different from a clinical point of view, which was evaluated together with a HF specialist. Hence, the value chosen for the analysis was $k = 4$.

C. Heart Failure Patient Subgroups

The clustering analysis resulted in four patient clusters, which were characterised in terms of comorbidities and demographic factors.

Patients from Cluster1 tend to be elder males. They show the highest number of ICD-9 codes and number of chronic diseases. It is also the cluster with the highest prevalence of almost all diseases, having a high percentage of patients with all the diseases analysed. The most common comorbidities are Anaemia (88.47%), CKD (74.52%), Hypertension (70.22%), Atrial Fibrillation (55.64%). There is also a high prevalence of ICM (49.43%) and Diabetes (30.42%). Patients belonging to this cluster also have very high values of NT-proBNP, usually related to a more severe state of HF. The median values for Sodium and Urea are above the reference values, which is in agreement with the high percentage of CKD and the median value for Hemoglobin is below the reference values, which is in agreement with the high percentage of patients with Anaemia.

Patients from Cluster2 are mostly elder women. The most common comorbidities in this cluster are Hypertension (84.53%), Atrial Fibrillation and Obesity (33.95%). Prevalence of Hypertension and Obesity is also higher in this cluster than in any other. Cluster2 patients present the highest medium value of BMI, which is related to the high percentage of patients with Obesity, and the lowest median value of NT-proBNP.

Cluster3 is the largest cluster (n=1231). Patients of Cluster3 are older men (59.46%) and women (40.54%) and have generally lower disease prevalence than the ones from Cluster2. The only disease where the prevalence is higher than in other clusters is Anaemia (99.35%). The median value of Hemoglobin is below the reference value in this cluster, which is in agreement with the high percentage of patients with Anaemia. The median value of NT-proBNP is the second highest when compared to other clusters.

Patients from Cluster4 are the youngest compared to other clusters and are predominantly female. These patients have the lowest number of diseases and the lowest prevalence of almost all diseases, except for Obesity (11.64%) and Cardiomyopathy (4.31%) where the prevalence is slightly higher than in Cluster3.

Figure 4 shows a tileplot of cluster-specific percentages of comorbidities that allows to easily find the most prevalent comorbidities in each cluster while comparing the prevalence of diseases among clusters. It is possible to confirm that Cluster1 has the highest percentage of almost all diseases, whereas Cluster4 can be seen as the low-burden cluster, considering the low percentage of other comorbidities. Cluster2 can be characterised by the high prevalence of Hypertension, Atrial Fibrillation and Obesity, and Cluster3 is the cluster with the highest percentage of patients with Anaemia.

Figure 5 shows the graph representation of each cluster's comorbidities. The representation can provide insights into cluster complexity. For example, it is possible to see that

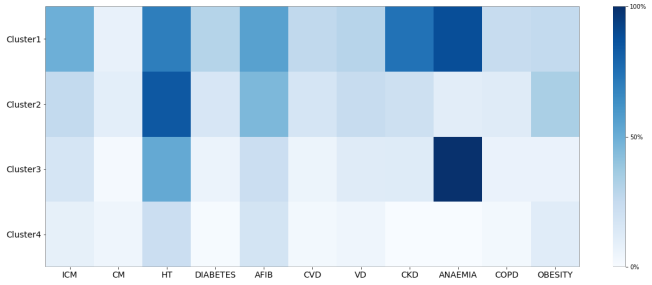
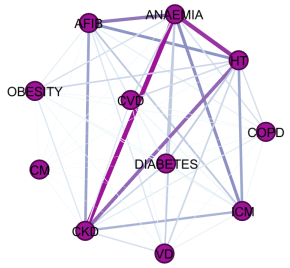
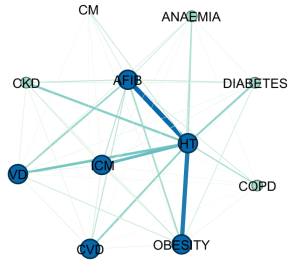


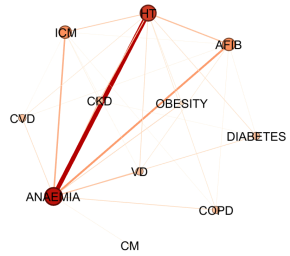
Fig. 4. Cluster-specific percentages of comorbidities. A darker colour indicates a higher percentage of the comorbidity in the cluster. ICM-Ischaemic Cardiomyopathy, HT-Hypertension, AFIB-Atrial Fibrillation, CVD-Cerebrovascular Disease, VD-Valvular Disease, CKD-Chronic Kidney Disease, COPD-Chronic Obstructive Pulmonary Disease



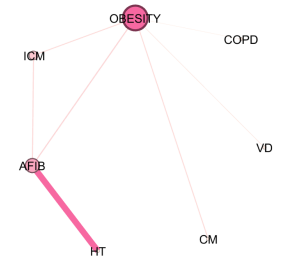
(a) Cluster1. Nodes = 11, edges = 55, average degree = 10.0, average clustering coefficient = 1



(b) Cluster2. Nodes = 11, edges = 49, average degree = 8.9, average clustering coefficient = 0.91



(c) Cluster3. Nodes = 11, edges = 28, average degree = 5.0, average clustering coefficient = 0.89



(d) Cluster4. Nodes = 7, edges = 7, average degree = 2.0, average clustering coefficient = 0.48

Fig. 5. Prevalence and graph representation of comorbidities used for clustering in the HF dataset. In the graph a node represents a disease and its size is proportional to the node degree. An edge represents a co-occurrence of two diseases and its width is proportional to the prevalence of the co-occurrence in the dataset. CM-Ischaemic Cardiomyopathy, HT-Hypertension, AFIB-Atrial Fibrillation, CVD-Cerebrovascular Disease, VD-Valvular Disease, CKD-Chronic Kidney Disease, COPD-Chronic Obstructive Pulmonary Disease

Cluster1 has the highest number of nodes and edges and the graph of Cluster4 has the lowest, which indicates that patients from Cluster1 have a higher complexity than patients from Cluster4. Figure 5 also illustrates the most common comorbidity associations in each cluster. Cluster1 has a high number of associations and a high number of strong associations between diseases. The fact that all nodes are of similar size means that every disease co-occurs at least once with almost all other diseases. The width of the edges is what allows us to understand which of these co-occurrences are more common. In Cluster1, these are Chronic Kidney Disease and Anaemia, and Hypertension and Anaemia. In Cluster2 there is a high number of patients with Obesity and Atrial Fibrillation, and Hypertension and Atrial Fibrillation. Cluster3 has a high number of patients with Hypertension and Anaemia, with some of these patients also having Atrial Fibrillation, Ischaemic Cardiomyopathy, and Valvular Disease. Looking at the graph from Cluster4, the most common association of diseases is Hypertension and Atrial Fibrillation. Obesity is also connected to several other diseases in this cluster, but with a lower co-occurrence. The average degree and average clustering coefficient are also helpful to quantify the clusters' complexity. The average degree indicates the average number of other diseases that are connected to one disease. We can observe that in Cluster1 the average degree is 10, which means all diseases are connected to each other. Cluster2 also has a high degree, 8.9, whereas Cluster3 and Cluster4 have much lower degrees, 5 and 2, respectively. The average clustering coefficient is a measure of density that indicates the degree to which nodes in a graph tend to cluster together. In this case, it can be interpreted as a measure of the tendency of diseases to co-occur. Also for this metric Cluster1 has the highest value and Cluster4 has the lowest (1 vs. 0.48), which suggests patients from Cluster1 have a higher order interactions of diseases than patients from Cluster4.

In terms of comorbidities, the identified clusters present similarities with the results reported in a study by Gulea et al. [23]. The study applied model-based clustering to 12 comorbidities of a cohort of HF patients and identified five clusters. The clusters were characterised by a different combination of comorbidities and socio-demographic factors, and named accordingly: low-burden, metabolic-vascular, Ischaemic, anaemic, and metabolic. The study found that patients in the metabolic-vascular cluster had the highest percentage of comorbidities and worst prognosis and that the patients in the low-burden cluster had the lowest percentage of comorbidities and best prognosis. It is also possible to identify a metabolic-vascular cluster, corresponding to Cluster1, an anaemic cluster, corresponding to Cluster3, and a low-burden cluster, corresponding to Cluster4. Cluster2, which can be identified as an obesity cluster, presents some similarities with the metabolic cluster found by Savarese et al. However, it lacks the high prevalence of Diabetes to be considered a metabolic cluster.

Prescriptions

Table V shows the results obtained for the mean number of prescriptions and percentages of specific medication groups in each cluster. The number of different medications is relative to the observation period for medical prescriptions (9 years and 8 months). Percentages of medication groups represent the percentage of patients that had at least one prescription of the relevant medication group during the period analysed.

Medication for HF treatment includes ACEis\ARBs, Beta-blockers, Diuretics, and Digoxins [38]. All of these groups

TABLE V
NUMBER OF MEDICATION AND MEDICATION GROUPS PREVALENCE PER CLUSTER AND IN THE ENTIRE DATASET. P-VALUES FOR THE COMPARISON OF THE CHARACTERISTICS ACROSS CLUSTERS. ACEI - ANGIOTENSIN–CONVERTING ENZYME INHIBITOR; MRA - ALDOSTERONE RECEPTOR ANTAGONISTS; DPP4I - DIPEPTIDYL PEPTIDASE-4 INHIBITOR

Medications	Cluster1	Cluster2	Cluster3	Cluster4	p-values
Patients with medication data	87.58	84.41	68.97	66.0	-
Avg prescriptions/year	6.6	4.15	2.53	2.00	-
Anticoagulants	42.98	37.76	26.5	19.58	<0.01
Statins	39.94	33.52	21.91	18.34	<0.01
Beta-Blockers	35.31	35.16	22.38	17.28	<0.01
Antiplatelets	34.88	24.62	21.08	10.76	<0.01
Inhalers Bronchodilator	32.27	23.53	16.49	13.76	<0.01
Diuretics	29.52	26.54	15.19	10.58	<0.01
ACEi \ARBs	40.96	28.32	21.29	17.11	<0.01
Hematinic factors	27.79	15.18	16.37	8.47	<0.01
Anticholinergics	23.59	15.73	11.9	8.64	<0.01
MRA	16.06	14.5	12.25	7.58	<0.01

are amongst the most prevalent medication groups in the dataset. Besides these medication groups, patients also have a high prevalence of Anticoagulants, Antiplatelets, Statins and Bronchodilators.

Cluster1 can be considered the cluster with the most severe stage of HF. It is thus expected that a high percentage of patients have prescriptions directly related to HF and also related to their comorbidities. The most common medication groups in each cluster are Anticoagulants (42.98%), ACEi\ARB (40.96 %), Statins (39.94%) and Beta-blockers (35.31%). There is also a high percentage of patients with prescriptions for Bronchodilators (32.27%), which are used for the treatment of COPD, and Diuretics (29.52%), commonly used to treat Hypertension or edema, often consequences of HF or CKD. In Cluster2 the most common medication group is Anticoagulants (37.76%), followed by Beta-blockers (35.16%), Statins (33.52%) and ACEi\ARB (28.32%). In this cluster, the most common comorbidities were Hypertension, Obesity, and Atrial Fibrillation, which are in agreement with the most common medication groups found. Patients from Cluster3 and Cluster4 have a lower percentage of all medication groups when compared to Cluster1 and Cluster2, with Cluster4 having the lowest percentage of all medication groups. It is possible to observe that only 16.37% of patients in Cluster4 have been prescribed Hematinic factors, while 99.35% of patients in this cluster were found to have Anaemia. Hematinic factors are often used to treat ferropenic Anaemia (Anaemia caused by iron deficiency). This is possibly due to an underdiagnosis of Anaemia or to Anaemia being due to other factors than iron deficiency.

Outcomes

We analysed the outcomes of hospitalisation, emergency admission, and mortality to understand if the clustering resulted in groups that had different risk associations with the outcomes considered. One important consideration for this analysis is that only events occurring during the observation period in HLL are considered, and so, it is possible that some patients experienced these events outside that time window or in other healthcare facilities.

TABLE VI
CHARACTERISATIONS OF OUTCOMES RELATED VARIABLES PER CLUSTER AND IN THE ENTIRE DATASET. CONTINUOUS VARIABLES ARE DESCRIBED AS MEDIAN (INTER-QUARTILE RANGE) AND CATEGORICAL VARIABLES AS %. P-VALUES FOR THE COMPARISON OF THE CHARACTERISTICS ACROSS CLUSTERS.

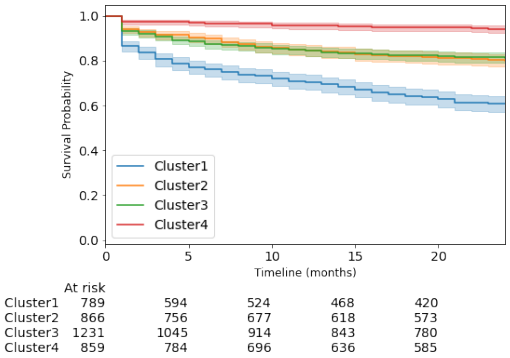
Characteristics	Cluster1	Cluster2	Cluster3	Cluster4	p-value
Number of Hospitalisations/year	0.2(0.1-0.4)	0.1(0.0-0.2)	0.1(0.0-0.2)	0.0(0.0-0.1)	<0.05
Hospitalisations within 1 year of HF diagnosis, %	35.23	23.78	27.86	21.18	<0.01
Hospitalisations within time period analysed, %	86.06	68.13	59.46	32.6	<0.01
Number of Emergency Admissions/year	0.6(0.2-1.2)	0.4(0.1-0.8)	0.2(0.1-0.5)	0.1(0.1-0.4)	<0.05
Emergency admissions within 1 year of HF diagnosis, %	52.47	44.34	39.48	36.20	<0.01
Emergency admissions within time period analysed, %	94.55	88.45	79.45	75.09	<0.01
Deceased, %	29.02	12.7	18.28	5.82	<0.01

The percentage of patients in the dataset with at least one hospitalisation is 60.91%, while 83.71% had at least one emergency admission (see Table VI). Cluster1 is the highest complexity cluster, which also translates into a higher percentage of patients that experienced hospitalisations and emergency admissions, and deceased patients. It is also the cluster with the highest number of hospitalisations and emergency admissions per year. Cluster2 has the second highest percentage of number of hospitalisations higher than one, while Cluster3 has the second highest percentage of deceased patients. Cluster4, the low-burden cluster, has the lowest percentages for all outcomes considered.

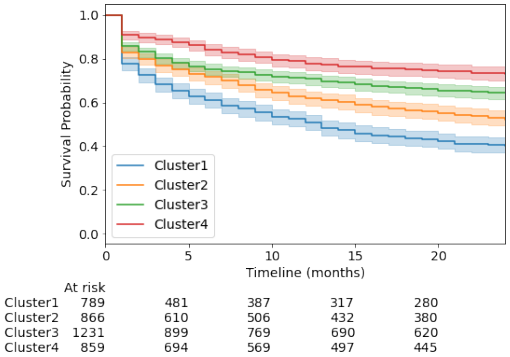
Survival Analysis

After having the clusters characterised in the domain of demographics, comorbidities, laboratory values, medical prescriptions, and outcomes prevalence. It was necessary to understand if there was an association between belonging to a cluster with a certain profile and the probability of experiencing a given clinical outcome. To do so, we conducted a Survival Analysis for the outcomes of hospitalisation and emergency for all patients in the dataset. Survival Analysis provides methods to determine how one group changes relative to another, in this case, how being in one cluster relative to being in another influences the outcome. mortality outcome due to the lack of longitudinal information.

Figure 6 shows univariate Kaplan-Meier curves for the outcomes of hospitalisation and emergency admission, stratified by clusters, for a time period of 2 years post HF diagnosis. Kaplan-Meier curves represent the survival probability of a population over time. In this case, each cluster represents a different population and the survival probability is represents not experiencing a hospitalisation or an emergency admission. At t_0 all clusters have a survival probability of 1, and as time moves forward, the survival probability goes down as a function of the number of patients experiencing the outcome



(a) Hospitalisation



(b) Emergency Admission

Fig. 6. Kaplan-Meier survival curves for the outcomes Hospitalisation and Emergency admission for each cluster (within 2-years after HF diagnosis)

in each cluster. Cluster1, as could be expected from the previous characterisation, appears as the highest risk cluster, with the lowest survival probability at all times. Contrarily, Cluster4 shows the highest survival probability during the period analysed. Cluster2 and Cluster3 have a similar survival function for the outcome hospitalisation. However, for the outcome of emergency admission Cluster3 has a lower survival probability than Cluster3. It is also interesting to note that the survival probability for the outcome emergency admission is, for all clusters, lower than for the outcome hospitalisation, suggesting that HF patients are more likely to have emergency admissions rather than being hospitalised. Differences between groups were tested using the log-rank test and the obtained p-value was <0.01 for both outcomes.

Besides Kaplan-Meier curves, Cox proportional hazard models were also computed to study how belonging to a certain cluster changes the rate of experiencing an outcome, with the possibility of taking into account other variables. One unadjusted and two multivariable adjusted Cox proportional hazards were computed (the results are displayed in Table VII). Model1 is unadjusted, Model2 is adjusted for baseline covariates Age and Gender, and Model 3 is adjusted for baseline covariates Age and Gender, and for NT-proBNP, often used as a risk marker for HF. The HR are computed in relation to the lowest risk cluster, Cluster4. All clusters show a higher risk for the outcomes of hospitalisation and

TABLE VII
RISK OF CLINICAL EVENTS HOSPITALISATION AND EMERGENCY ADMISSION COMPARED WITH CLUSTER4 (LOWEST RISK). HAZARD RATIOS AND 95% CONFIDENCE INTERVALS COMPUTED USING COX REGRESSION. ADJUSTED FOR AGE, GENDER AND NT-PROBNP (MODEL3). MODEL1 ADJUSTED FOR AGE AND GENDER, MODEL 2 ADJUSTED FOR AGE, GENDER AND NT-PROBNP.

	Cluster1	Cluster2	Cluster3	Cluster4	p-value
Model1, HR (95% CI)					
Hospitalisation	5.86 (4.80-7.15)	2.82 (2.29-3.48)	2.43 (1.98-2.98)	1	<0.05
Emergency Admission	2.73 (2.38-3.14)	2.00 (0.84-2.31)	1.29 (1.13-1.49)	1	<0.05
Model2, HR (95% CI)					
Hospitalisation	4.90 (3.97-6.05)	2.57 (2.08-3.18)	2.10 (1.70-2.60)	1	<0.05
Emergency Admission	2.60 (2.24-3.02)	1.92 (1.66-2.21)	1.24 - (1.07-1.44)	1	<0.05
Model3, HR (95% CI)					
Hospitalisation	4.73 (3.83-5.84)	2.58 (2.09-3.19)	2.10 (1.70-2.60)	1	<0.05
Emergency Admission	2.58 (2.22-2.99)	1.92 (1.66-2.21)	1.24 (1.07-1.44)	1	<0.05

emergency admission. This risk is the highest for Cluster1 and the lowest for Cluster3. In Model1, HR for hospitalisation ranged from 5.86 (4.80 - 7.15) for Cluster1 to 2.43 (1.98 - 2.98) for Cluster3, when compared with Cluster4. For the outcome hospitalisation, HR ranged from 2.73 (2.38 - 3.14) for Cluster1 to 2.00 (0.84 - 2.31) and 1.29 (1.13 - 1.49) for Cluster3. Despite having slightly lower values, differences in the HR of both outcomes remained significant when adjusting for Age and Gender, in Model2, and also when adjusting for NT-proBNP, in Model3.

The results obtained in this section are in agreement with the cluster profiles obtained from previous sections. Cluster1 has the highest risk profile, with older patients having a high number of comorbidities. Cluster2, a slightly younger cluster, with a high prevalence of AFIB, Obesity and HT, has the second highest risk for hospitalisation and emergency admissions. Cluster3 patients, with a high prevalence of Anaemia, show moderate risk for the outcomes considered. The lowest risk cluster is Cluster4, whose patients are younger and have the least prevalence of comorbidities. The results obtained show that different clusters are associated with different levels of risk for the outcomes considered, and thus cluster membership can be used for risk stratification

V. CONCLUSIONS

In this work we developed a workflow to identify and phenotype subgroups of HF patients with multimorbidity, using real-world EHR data from HLL. Four clusters were identified, which present significant differences in clinical and demographic characteristics, along with different prevalence of comorbidities. We were able to trace a clinical profile for each cluster. Medical prescriptions also presented differences between the patient subgroups. Additionally, the survival analysis showed that each HF subgroup is associated with different levels of risk for the outcomes of hospitalisation and emergency admission. The results from the survival analysis were in agreement with the profile traced for each cluster. Clusters with higher complexity (higher number of comorbidities and medical prescriptions) had a worse prognosis and patients with a lower complexity had a better prognosis.

Future work should focus on cross-validation with an external cohort, which can be done by replicating this approach in a dataset from a different hospital or healthcare facility. This

will be key to determining whether the obtained HF clusters and their phenotypes found in this study are replicable in other populations and EHR systems.

Additionally, future studies could benefit from integrating the EF parameter, either as a feature used for clustering or to analyse if patient subgroups resulting from the clustering have similar EF values.

HF is a highly heterogeneous syndrome. In this study, it was possible to observe this with real patient data containing information from different sources within a hospital.

REFERENCES

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

[2] Caroline Bähler et al. "Multimorbidity, Health Care Utilization and Costs in an Elderly Community-dwelling Population: A Claims Data Based Observational Study". In: *BMC health services research* 15 (Jan. 2015), p. 23. DOI: 10.1186/s12913-015-0698-2.

[3] Francesca Colombo, Manuel Goñi, and Christoph Christoph. "Addressing Multimorbidity to Improve Healthcare and Economic Sustainability". In: *Journal of Comorbidity* 6 (Feb. 2016), pp. 21–27. DOI: 10.15256/joc.2016.6.74.

[4] Juan Banda et al. "Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models". In: *Annual Review of Biomedical Data Science* 1 (July 2018). DOI: 10.1146/annurev-biodatasci-080917-013315.

[5] National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. Jan. 2012, pp. 1–128. DOI: 10.17226/13284.

[6] Chaitanya Shivade et al. "A Review of Approaches to Identifying Patient Phenotype Cohorts Using Electronic Health Records". In: *Journal of the American Medical Informatics Association : JAMIA* 21 (Nov. 2014). DOI: 10.1136/amiajnl-2013-001935.

[7] Gary S. Francis, Rebecca Cogswell, and Thenappan Thenappan. "The Heterogeneity of Heart Failure". In: *Journal of the American College of Cardiology* 64.17 (2014), pp. 1775–1776. DOI: 10.1016/j.jacc.2014.07.978.

[8] Shireen Mirza and Roberto Benzo. "Chronic Obstructive Pulmonary Disease Phenotypes: Implications for Care". In: *Mayo Clinic Proceedings* 92 (July 2017), pp. 1104–1112. DOI: 10.1016/j.mayocp.2017.03.020.

[9] Mary Ann Thenganatt and Joseph Jankovic. "Parkinson Disease Subtypes". In: *JAMA Neurology* 71.4 (Apr. 2014), pp. 499–504. ISSN: 2168-6149. DOI: 10.1001/jamaneurol.2013.6233.

[10] Tariq Ahmad et al. "Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients". In: *Journal of the American Heart Association* 7 (Apr. 2018), e008081. DOI: 10.1161/JAHA.117.008081.

[11] Chantal Raherison et al. "Comorbidities and COPD severity in a clinic-based cohort". In: *BMC Pulmonary Medicine* 18 (July 2018). DOI: 10.1186/s12890-018-0684-7.

[12] Amy Groenewegen et al. "Epidemiology of heart failure". In: *European Journal of Heart Failure* 22 (June 2020). DOI: 10.1002/ejhf.1858.

[13] Gianluigi Savarese and Lars Lund. "Global Public Health Burden of Heart Failure". In: *Cardiac Failure Review* 03 (Apr. 2017), p. 7. DOI: 10.15420/cfr.2016:25:2.

[14] Alexander Foss, Marianthi Markatou, and Bonnie Ray. *Distance Metrics and Clustering Methods for Mixed-Type Data*. Apr. 2018.

[15] John Gower. "A General Coefficient of Similarity and Some of Its Properties". In: *Biometrics* 27 (Dec. 1971), pp. 857–871. DOI: 10.2307/2528823.

[16] Shatrunjai Singh et al. "Agglomerative Hierarchical Clustering Analysis of co/multi-morbidities". In: (July 2018).

[17] Francois Husson, Julie Josse, and Sébastien Lê. "FactoMineR: An R Package for Multivariate Analysis". In: *Journal of Statistical Software* 25 (Mar. 2008). DOI: 10.18637/jss.v025.i01.

[18] Davide Vetrano et al. "Twelve-year clinical trajectories of multimorbidity in a population of older adults". In: *Nature Communications* 11 (June 2020), p. 3223. DOI: 10.1038/s41467-020-16780-x.

[19] Concepcion Violan et al. "Multimorbidity patterns with K-means nonhierarchical cluster analysis". In: *BMC Family Practice* 19 (July 2018). DOI: 10.1186/s12875-018-0790-x.

[20] Tariq Ahmad et al. "Clinical Implications of Chronic Heart Failure Phenotypes Defined by Cluster Analysis". In: *Journal of the American College of Cardiology* 64 (Oct. 2014), pp. 1765–1774. DOI: 10.1016\/j.jacc.2014.07.979.

[21] Clyde W. Yancy et al. "2013 ACCF/AHA Guideline for the Management of Heart Failure". In: *Circulation* 128.16 (2013), e240–e327. DOI: 10.1161/CIR.0b013e31829e8776.

[22] Biykem Bozkurt, Andrew Coats, and Hiroyuki Tsutsui. "Universal Definition and Classification of Heart Failure". In: *Journal of cardiac failure* 27 (Feb. 2021). DOI: 10.1016/j.cardfail.2021.01.022.

[23] Claudia Gulea, Rosita Zakeri, and Jennifer Quint. "Model-based comorbidity clusters in patients with heart failure: association with clinical outcomes and healthcare utilization". In: *BMC Medicine* 19 (Jan. 2021). DOI: 10.1186/s12916-020-01881-7.

[24] Taane Clark et al. "Survival Analysis Part I: Basic Concepts and First Analyses". In: *British Journal of Cancer* 89 (Aug. 2003), pp. 232–8.

[25] Viv Bewick, Liz Cheek, and Jonathan Ball. "Statistics review 12: Survival analysis". In: *Critical care (London, England)* 8 (Nov. 2004), pp. 389–94. DOI: 10.1186/cc2955.

[26] C. Cramer et al. *Network Literacy: Essential Concepts and Core Ideas*. Mar. 2015.

[27] Derek L. Hansen, Ben Shneiderman, and Marc A. Smith. "Chapter 3 - Social Network Analysis: Measuring, Mapping, and Modeling Collections of Connections". In: *Analyzing Social Media Networks with NodeXL*. Ed. by Derek L. Hansen, Ben Shneiderman, and Marc A. Smith. Boston: Morgan Kaufmann, 2011, pp. 31–50. ISBN: 978-0-12-382229-1. DOI: https://doi.org/10.1016/B978-0-12-382229-1.00003-5.

[28] Cesar Hidalgo et al. "A Dynamic Network Approach for the Study of Human". In: *PLoS computational biology* 5 (May 2009), e1000353. DOI: 10.1371/journal.pcbi.1000353.

[29] Akbar Waljee et al. "Comparison of imputation methods for missing laboratory data in medicine". In: *BMJ open* 3 (Aug. 2013). DOI: 10.1136/bmjopen-2013-002847.

[30] Melissa Azur et al. "Multiple Imputation by Chained Equations: What is it and how does it work?" In: *International journal of methods in psychiatric research* 20 (Mar. 2011), pp. 40–9. DOI: 10.1002/mpr.329.

[31] Gizem Yilmaz and Hira Shaikh. "Normochromic Normocytic Anemia". In: (Dec. 2020).

[32] OECD/WHO. *Overweight and obesity*. 2020. DOI: /https://doi.org/10.1787/a47d0cd2-en.

[33] Frank Nielsen. "Hierarchical Clustering". In: *Introduction to HPC with MPI for Data Science* (2016), pp. 195–211. DOI: 10.1007/978-3-319-21903-5_8.

[34] A.D. Gordon. *Classification*. 1999.

[35] Peter Rousseeuw. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65". In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[36] T. Caliński and J Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics* 3.1 (1974), pp. 1–27. DOI: 10.1080/03610927408827101.

[37] David L. Davies and Donald W. Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

[38] European Society of Cardiology. "Acute and Chronic Heart Failure Guidelines. ESC Clinical Practice Guidelines." In: (2016).

[39] Sanjiv Shah et al. "Phenomapping for Novel Classification of Heart Failure With Preserved Ejection Fraction". In: *Circulation* 131 (Nov. 2014). DOI: 10.1161/CIRCULATIONAHA.114.010637.