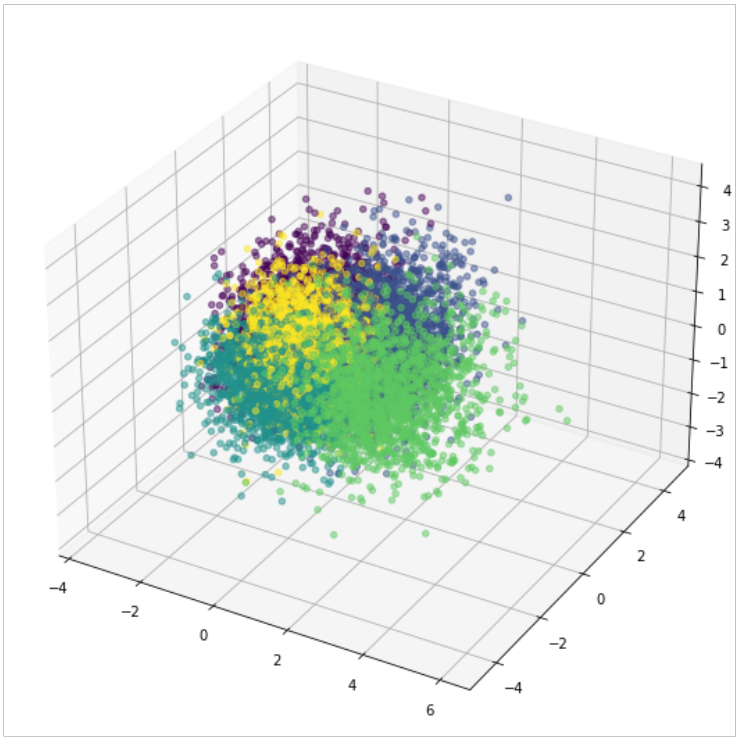

Data Driven Heart Failure Patient Segmentation

Identification of Underlying Patient Phenotypes

Linn Hellberg
Sushanthika Gnanasekaran

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Data Driven Heart Failure Patient Segmentation

Identification of Underlying Patient Phenotypes

Linn Hellberg
Sushanthika Gnanasekaran

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Division of Biomedical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2025

Data Driven Heart Failure Patient Segmentation
Identification of Underlying Patient Phenotypes
Linn Hellberg
Sushanthika Gnanasekaran

Supervisor: Ruben Buendia, AstraZeneca
Examiner: Hana Dobicek Trefna, Department of Electrical Engineering

Master's Thesis 2025
Department of Electrical Engineering
Division of Biomedical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Principal component analysis of clusters obtained from latent class analysis.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Data Driven Heart Failure Patient Segmentation
Identifying Underlying Patient Phenotypes to Aid in Aetiological Discovery
Linn Hellberg, Sushanthika Gnanasekaran
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Heart failure is a heterogeneous syndrome where the underlying aetiology remains uncertain. The most common sub-classifications of heart failure are based on left ventricular ejection fraction, forming three subgroups: reduced (HFrEF), mildly reduced (HFmrEF), and preserved (HFpEF). While the development of targeted therapies holds promise for improving outcomes across the heart failure spectrum, current sub-classifications have not yet enabled precision medicine, as they fail to fully capture the underlying pathophysiological mechanisms.

Recent efforts have focused on HFpEF, and thus, patient segmentation in heart failure remains suboptimal. Based on data from four clinical trials in heart failure with a total of 11,140 patients, this project aimed to identify clinically relevant clusters across the full spectrum of ejection fraction. Twelve numerical and categorical patient characteristics were used as input covariates in latent class analysis to model underlying distributions, identifying four distinct clusters:

- Cluster 1: Old atrial fibrillation no myocardial infarction (HFpEF, 29.1% of patients)
- Cluster 2: Male high NT-ProBNP with myocardial infarction (HFrEF, 29.2% of patients)
- Cluster 3: Obese diabetic (HFpEF, 22.0% of patients)
- Cluster 4: Young male, with good kidney function (HFrEF, 19.8% of patients)

The identified patient sub-groups are clinically meaningful and associated with significantly different hard outcomes, such as cardiovascular and all-cause mortality. A sensitivity analysis using a k-prototypes clustering algorithm for mixed data derived clusters that corresponded to most characteristics of the HFrEF groups. However, cluster separation was relatively low across both latent class analysis and k-prototypes. Further work, such as data-driven feature selection, could improve cluster quality and separation.

Keywords: Phenotyping, Heart failure, Latent class analysis, Clustering, Unsupervised machine learning.

Acknowledgements

We would like to sincerely thank Ruben Buendia Lopez, our supervisor at AstraZeneca, for his invaluable guidance, encouragement, and for being the one who brought this project to life. We wholeheartedly appreciate our examiner, Hana Dobicek Trefna, Department of Electrical Engineering at Chalmers, for her academic guidance and supervision during the project. We also extend our thanks to our manager, Martin Karpefors, and the rest of our team at Data Science in Late CVRM at AstraZeneca, for their collaborative support, input along the way, and for making us feel welcome. We are thankful to Dr. Ola Vedin, Global Clinical Head, Late Clinical CVRM, AstraZeneca, for taking the time to help us interpret and validate the clinical relevance of our results. Finally, we would like to express our gratitude to AstraZeneca for the opportunity to work on this project. This thesis marks the culmination of our degree, and we would like to thank everyone who played a part in shaping our experience throughout the past years.

Linn Hellberg, Sushanthika Gnanasekaran, Gothenburg, 06 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

AF	Atrial Fibrillation
AIC	Akaike Information Criterion
BIC	Bayesian information criterion
BLRT	Bootstrapped Likelihood Ratio Test
BMI	Body Mass Index
BPM	Beats per Minute
CKD	Chronic Kidney dDisease
CO	Cardiac Output
COPD	Chronic Obstructive Pulmonary Disease
CV	Cardiovascular
eGFR	Estimated Glomerular Filtration rate
EM	Expectation Maximization
FAS	Full Analysis Set
GMM	Gaussian Mixture Model
HFmrEF	Heart Failure with mildly reduced Ejection Fraction
HFpEF	Heart Failure with preserved Ejection Fraction
HFrEF	Heart Failure with reduced Ejection Fraction
FMM	Finite Mixture Models
HbA1c	Hemoglobin A1c
HF	Heart Failure
KCCQ	Kansas City Cardiomyopathy Questionnaire
LCA	Latent Class Analysis
LPA	Latent Profile Analysis
LVEF	Left Ventricular Ejection Fraction
MI	Myocardial Infarction
NT-proBNP	N-terminal pro-B-type Natriuretic Peptide
NYHA	New York Heart Association
PCA	Principal Component Analysis
SS	Silhouette Score
T2DM	Type 2 Diabetes Mellitus
TTE	Time To Event
WCSS	Within Cluster Sum of Squares

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Indices

i, j	Indices for data points
k	Index over K clusters
l	Index for F number of features
p	Index over P numerical features
q	Index over Q categorical features
r	Index over R ordinal categories

Sets

$\mathbf{x_i}$	Vector of observed continous variables for individual i
$\mathbf{u_i}$	Vector of observed categorical variables for individual i

Parameters

π_k	Mixing proportion for class k
μ_k	Mean of class k
σ_k^2	Variance of class k
γ	K-prototypes Categorical weighting
ω_{ijk}	Gower's Distance feature weights
K	Number of clusters
g	Number of model parameters for BIC

Variables

s_{ijl}	Similarity between i and j for feature l
D_{ijl}	Gower's distance between i and j over feature l
SS	Silhouette score
BIC	Bayesian information criterion
H	Entropy

Contents

List of Figures

List of Tables

1
Introduction

Heart failure (HF) is a progressive, heterogeneous clinical syndrome, traditionally characterized by the heart's inability to maintain sufficient cardiac output (CO) to meet the metabolic needs of the body [1], [2]. This is often caused by abnormalities in the structure or function of the cardiac muscle [3]. Heart failure is a syndrome rather than a single disease, evident from a collection of symptoms, with varied aetiology and overlapping clinical presentation. According to a systematic analysis for the Global Burden of Disease Study, the adult population has a heart failure prevalence of 1-3% corresponding to an estimate of 64.3 million HF patients worldwide [2].

Decreased cardiac function manifests as fatigue, edema, and dyspnea at low exertion [3]. Despite advances in healthcare, heart failure continues to be associated with high symptom burden and mortality, particularly in older patients. These symptoms reduce functional capacity and impair quality of life. Additionally, the syndrome has a range of comorbidities that play a crucial role in influencing the disease prognosis [3].

Heart failure is commonly classified using Left Ventricular Ejection Fraction (LVEF) into three categories: heart failure with reduced ejection fraction (HFrEF, LVEF $\leq 40\%$), mildly reduced ejection fraction (HFmrEF, LVEF 41 - 49%), and preserved ejection fraction (HFpEF, LVEF $\geq 50\%$) [3]. These categories reflect distinct underlying pathophysiologies. HFrEF is defined as a systolic dysfunction, caused by decreased contractility of the myocardium, resulting in reduced output from the left ventricle [4]. In contrast, HFpEF is characterized by diastolic dysfunction caused by impaired relaxation, which reduces ventricular filling and cardiac output, while HFmrEF has properties from both conditions.

While pharmacotherapy for HFrEF is effective, HFpEF has a highly heterogeneous aetiology with many clinical phenotypes, making it difficult to identify targeted therapies [3]. However, mortality in HFrEF patients remains high, despite advances in treatment that reduces hospitalization rates and improves outcomes. Diagnosis of HFmrEF is complicated as it shares characteristics of HFrEF and HFpEF, resulting in no strong treatment recommendations. It has further been suggested that categorizing patients only by LVEF hinders phenotyping and development of targeted treatments, as it is currently only effective for HFrEF [5].

The heterogeneity might be better understood through discovering new heart failure phenotypes [6], which in turn could allow for refined clinical trial inclusion criteria leading to phenotype-specific trials and driving the development of targeted

therapies. Grouping patients by disease phenotype may facilitate the discovery of phenotype-specific treatments [7]. Applying machine learning across the full spectrum of LVEF could allow for identification of novel HF subtypes [5], [6].

Clustering is one method that uses machine learning to group similar patients together and dissimilar patients apart [8]. Recent efforts to phenotype heart failure regularly find between 2 and 6 patient clusters [9], often solely focusing on HFpEF patients [6]. However, many papers cluster on large feature spaces [10]–[17], complicating interpretability and communication of the identified clusters, partly as they are difficult to visualize [18], but also as the clinical usability of such clusters has been questioned.

While recent studies aim to uncover novel heart failure subtypes, further research is needed for validation of models and subsequent clusters [6]. Many papers attempt to find optimal clusters and repeat the process on a validation set to assess reproducibility, but do not report cluster quality metrics [11]–[13], [16], [19], [20]. Although reproducibility partly validates the clusters, it does not necessarily communicate intra-cluster cohesion or inter-cluster separation, which is crucial for future clinical usability [21].

1.1 Aim

To address the aforementioned limitations, this thesis proposes to apply clustering to heart failure trial datasets across the full ejection spectrum to aid in the understanding of this complex, heterogeneous syndrome by hypothesizing its pathophysiology as a proxy of the aetiology. This will be presented as clusters of clinical patient data.

1.2 Objectives

The main objective of this thesis is to phenotype heart failure patients into specific subgroups that can be clinically useful. The following sub-objectives will help attain the aim of the project:

1. *Choose* suitable features and clustering techniques.
2. *Cluster* patients using clinical trial data with different methods.
3. *Evaluate* the prognosis per cluster and compare it to phenotypes in the current literature.
4. *Hypothesize* about cluster-specific pathophysiology.

1.3 Limitations

This thesis work was conducted during the spring term of 2025, with a clearly defined scope of the project. The analysis was based on the clinical trial data provided

and carefully limited by AstraZeneca. Clinical datasets often do not follow patients' pre-disease incidence. Instead, it had to suffice to infer pathophysiology from phenotyping and evaluate outcomes in different clusters from a cross-sectional cohort. Phenotyping could be a step towards precision medicine, thus furthering the research in phenotype-based targeted heart failure treatments. However, recommending improved clinical-trial populations was outside of the scope of this project. Further, the project was limited to packages implemented in Python.

1.4 Use of generative AI

This thesis employs AI tools as permitted by Chalmers under *'Regulations for the use of AI tools in thesis work'* [22]. Having consulted with our supervisor and examiner, ChatGPT [23] was used for this purpose.

The extent of use of AI was strictly limited to:

1. Used in addition to literature to understand difficult topics or make inferences from literature.
2. Identifying opportunities for compacting text and clarifying language.
3. Identifying and correcting grammatical inconsistencies in the text.

2
Theory

This chapter presents the theory behind heart failure as a complex syndrome and clustering techniques that lay the foundation for the work conducted in this project.

2.1 Heart Failure - A Complex Syndrome

To understand the challenges posed by the heterogeneity of heart failure, it is important to recognize the disease as a syndrome with overlapping comorbidities and risk factors such as: diabetes, obesity, hypertension, Atrial Fibrillation (AF), Myocardial Infarction (MI), Chronic Kidney Disease (CKD) and Chronic Obstructive Pulmonary Disease (COPD) [3].

Chronic kidney disease is defined by decreased kidney function over three months, with an estimated glomerular filtration rate (eGFR) $\leq 60ml/min/1.73m^2$ or presence of albuminuria [24]. The condition causes cardiac oxidative stress and fibrosis partly through hormonal and inflammatory processes, resulting in concentric cardiac remodeling, see Figure 2.1 [24]. This leads to decreased filling volume and diastolic dysfunction, resulting in HFpEF [25]. While CKD is more common in HFpEF, it is also comorbid with HFrEF. Many pathophysiological mechanisms in HF and CKD interact and worsen both conditions [24]. As they share comorbidities and risk factors, diagnosis is difficult due to overlapping clinical presentations.

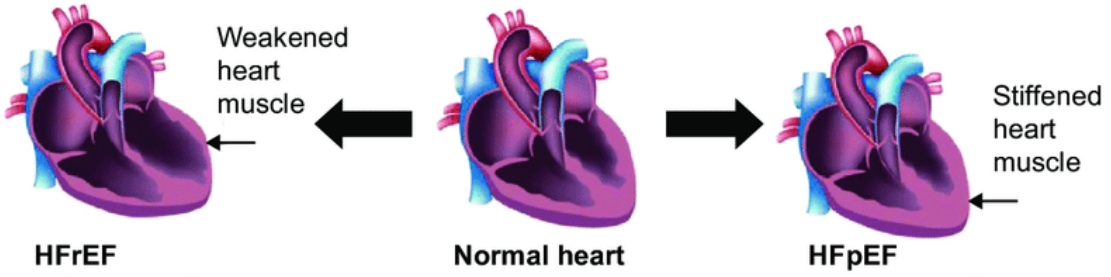

Figure 2.1: Concentric cardiac remodelling leading to HFpEF or eccentric cardiac remodelling leading to HFrEF [26].

Diabetes represents another comorbidity with HF [27]. One of the diagnostic criteria for type 2 diabetes mellitus (T2DM) is glycated hemoglobin levels (HbA1c) $\geq 6.5\%$ or fasting plasma glucose $\geq 7mmol/l$, representing long-term and current

blood glucose levels, respectively. The resulting hyperglycemia acts through complex pathways and can result in an inflammatory response, fibrotic myocardium, stiffness, and diastolic dysfunction, thus increasing the risk of, or worsening existing HF [28]. Although HFpEF is more common and severe for this comorbidity, HFrEF is also prevalent. Obesity, which is defined by a body mass index (BMI) $\geq 30 kg/m^2$, is a risk factor for both T2DM and HF [29]. As obesity causes increased blood volume, subsequent cardiac volume overload and elevated CO, it leads to eccentric cardiac remodeling, see Figure 2.1. Although being more prevalent for HFpEF, it could lead to HFrEF depending on existing comorbidities such as T2DM.

Coronary Artery Disease (CAD) is the most common cause for HF in industrialized countries [3]. It is caused by blockage of the coronary arteries, mainly due to atherosclerosis [30]. The condition results in hypoxia, fibrosis or loss of functioning myocardium, causing eccentric cardiac remodeling, and systolic dysfunction leading to HFrEF, see Figure 2.1 [31]. Although more common for HFrEF, ischemia can cause fibrosis, concentric remodeling that impairs relaxation, and lead to HFpEF.

Heart failure can further cause acute cardiac events such as MI, obstructing blood vessels, causing hypoxia, death of myocardium, loss of contractility, and exacerbating HFrEF. Another comorbidity is AF, which can precede or follow HF, both sharing common comorbidities [3]. The condition is defined as arrhythmia of the atrium, which causes decreased CO through impairment of ventricular filling, thus leading to HFpEF [32]. However, prolonged AF with resulting tachycardia can cause cardiomyopathy, leading to HFrEF [33].

Because of multicomorbidities, it is unclear where one condition ends and another begins. One example being the interactions between AF and HF, where the diagnosis of AF changes the diagnostic criteria needed for HF [34]. For a diagnosis of HF in a patient without AF, it is required that a marker for intracardiac pressure called N-terminal pro-brain natriuretic peptide (NT-ProBNP) $\geq 125 pg/mL$ [3]. However, in patients with pre-existing AF, NT-ProBNP is further increased, complicating the diagnosis [34]. Another example of a comorbidity altering the presentation of HF is when it co-exists with obesity, as NT-ProBNP is broken down by an enzyme released from fatty tissue [29].

Additionally, comorbidities accumulate with age where long disease duration might worsen conditions and alter clinical presentation over time [35]. One such example being patients with long duration of obesity and HF that lose weight due to multicomorbidity and severe illness, which effectively change the clinical presentation [29], which is one reason why symptom burden is an important marker. This is often evaluated through New York Heart Association (NYHA) classification or Kansas City Cardiomyopathy Questionnaire (KCCQ) [36]. In general, KCCQ more accurately reflects HF mortality and could assess disease burden as comorbidities accumulate or worsen over time [37].

2.2 Clustering

Clustering is a form of unsupervised machine learning that groups similar data points and separates dissimilar ones into different clusters. As outlined in Data Clustering: Algorithms and Applications [8], various techniques have been developed, each suited to different data structures and objectives.

One type of clustering is called partitional or prototype-based, which assigns data points into a predefined number of clusters by optimizing an objective function based on the distance, or similarity, between a data point and its cluster center, represented by a centroid or medoid [8]. These methods rely on distance metrics, the most common being euclidean distance for numerical data. The centroid is the average of numerical data points in a cluster, while a medoid is a data point that represents the center. K-means is the simplest algorithm which minimizes the squared euclidean distance to each cluster's centroid, while k-medoids minimizes the absolute error between the data points and their respective medoids. Additionally, k-modes is the categorical counterpart, minimizing dissimilarity based on mismatches between the data point and the most common category in that cluster. Partitioning methods such as these often employ hard clustering by assigning data points to exactly one cluster and iteratively refining the objective function.

In contrast, hierarchical clustering is a deterministic, distance based method that organizes clusters and their sub-clusters in a hierarchy [8]. In the case of agglomerative clustering, it starts with every data point assigned as its own cluster and recursively merges them based on a similarity measure until all points are assigned to one single cluster. Hierarchical clustering outputs a dendrogram, a tree-like diagram that visualizes data points organized into a hierarchy of clusters [8].

Another class of clustering is model-based methods, such as Gaussian Mixture Models (GMM), which assume that data are generated from underlying and overlapping probability distributions [8]. These approaches enable soft clustering, where data points may belong to multiple clusters. The algorithm assigns each point a probability of belonging to each cluster, allowing for a more granular interpretation of cluster membership. Another type of method is density-based, which detect clusters based on regions of high data density, making them effective for non-convex shapes and outlier handling.

2.2.1 Preprocessing

Real-world data, often obtained from varied resources, is inconsistent, noisy, and has missing values [21]. The efficacy of clustering algorithms is data-dependent, where higher data quality and structure improve the robustness of the clustering outcomes. Consequently, different machine learning models require data in various formats. For instance, clustering algorithms, specifically based on distances, are often unsuitable for mixed data and are outlier sensitive [38]. Preprocessing techniques transform raw data by eliminating inconsistencies, reducing noise, and

dimensional redundancy, making the data suitable for clustering.

Data cleaning is the first step in preprocessing, where missing values, noise, and outliers are identified and addressed [21]. Missingness and inconsistencies in clinical data are commonly caused by incorrect data collection or entry, equipment failures, etc. Many clustering algorithms, such as k-means, are incompatible with data containing missing values. The standard solution is to include listwise or pairwise deletion, as well as imputation by mean substitution, k-nearest neighbor, or multiple imputation by chained equations [39]. When the extent of missingness is less than 5 percent and presumed to be missing at random, bias introduced by imputations is likely to be insignificant [40]. However, certain state-of-the-art model-based algorithms, such as Latent Class Analysis (LCA), can handle missing values internally [41]. Hence, the choice of using a complete or imputed dataset depends on the algorithm in use.

The cluster quality in partition-based algorithms is influenced by outliers and duplicate data entries [21]. For example, for many partition-based methods, outliers might affect the averages that is used to calculate the cluster centroids. Outliers can be identified by visual or statistical methods and be clipped or removed to prevent calculating misleading cluster centers, which affect cluster separation. Additionally, duplicate data points introduce redundancy which distort similarity or distance measures [21]. Therefore, they are identified and deleted to maintain data integrity.

The scale of different variables can affect the analysis and clustering of the data [21]. Variables with larger range outweigh the variables with a smaller range, leading to biased results [38]. One solution is normalization which aim to give all variables equal weight to overcome this bias. Min-max normalization rescales the data into a fixed range [0, 1] [21]. Alternatively, Z-score normalization centers the data around the mean of 0 with a standard deviation of 1.

Clustering performance is also sensitive to the shape and distribution of the data [21]. Distribution based methods such as GMMs, which often assume that data follows a gaussian or symmetric distribution [8], an assumption frequently violated in real-world data. A common solution is to apply a log transformation to non-normal and right-skewed data, which compresses extreme values and reduces skewness. Other transformations like Box-Cox or Yeo-Johnson can further normalize both right- and left-skewed data [42].

In high-dimensional settings, clustering algorithms often suffer the curse of dimensionality, where distance-based similarity measures are less effective as distances between data points increase and become seemingly uniform [43], [44]. Feature extraction methods such as Principal Component Analysis (PCA) is one of the most established techniques for dimensionality reduction [8]. It projects high-dimensional data into lower dimensional subspaces by constructing the linear combinations of features called principal components [45]. These principal components are chosen so that they capture the most variance of the dataset, thereby preserving its infor-

mation. Non-linear projection methods such as Uniform Manifold Approximation and Projection are frequently employed for visualization purposes [46].

Feature selection is another method that reduces dimensionality by isolating informative attributes, while removing redundant or irrelevant features, that obscure the true cluster structure [21]. Filtering techniques are a method of feature selection that remove redundancies [47]. Redundant covariates can be detected, for example through correlation analysis using Pearson's coefficient for continuous variables, and the Chi-square test for nominal attributes [21]. Furthermore, for clinical datasets, domain knowledge is critical for feature selection to prioritize clinically significant and biologically relevant variables [47].

2.2.1.1 Categorical Encoding

When clustering mixed data such as numerical, categorical and ordinal variables, appropriate transformations must be applied. To enable distance-based computations, categorical variables can be encoded as numerical [48]. For ordinal variables, categories with low frequencies are commonly merged with adjacent categories [41]. However, encoding unordered categories as ordinal variables misrepresents distances between data points. One-hot encoding is another solution, which creates one binary feature for each category in a variable [8]. For variables with many categories, this creates a high dimensional feature space, causing the variable to contribute more to distance calculations and leading to inflated importance [49].

2.2.2 Clustering with Mixed Data

Many clustering methods are designed to handle either numerical or categorical data, but few natively support mixed datasets [48], thus limiting their applications as many real-world datasets include a combination of data types. While model-based methods often handle mixed data natively, distance-based methods such as k-means, commonly require alternative solutions [50].

2.2.2.1 Gower's Distance

Gower's distance is an alternative which allows for similarity calculations with mixed data, that computes the numerical and categorical distances separately [50]. For numerical variables x it calculates a similarity between data points i and j for variable l, as shown in Equation 2.1:

$$s_{ijl} = \frac{|x_{il} - x_{jl}|}{\max(x_l) - \min(x_l)} \tag{2.1}$$

For categorical variables, the similarity is based on matches between data points i and j for variable l, as described in Equation 2.2:

$$s_{ijl} = \begin{cases} 1 & \text{if } x_{il} = x_{jl} \\ 0 & \text{if } x_{il} \neq x_{jl} \end{cases} \tag{2.2}$$

Finally, Gower's distance sums the distances for the features p, both categorical and numerical covariates with optional weighting ω_{ijl} of each feature, as in Equation 2.3.

$$D_{ijl} = 1 - \frac{\sum_{l=1}^{p} s_{ijl} \cdot \omega_{ijl}}{\sum_{l=1}^{p} \omega_{ijl}} \tag{2.3}$$

Weights are typically set to $\omega_{ijl} = 1$, resulting in a heavy categorical influence. This is because the categorical similarity often reaches values of 0 or 1, while numerical values are normalized and only have a similarity of zero when the maximum and minimum values are compared. To resolve this issue, feature weights are commonly used, however often set manually [51]. As many mixed-methods suffer this problem, options for calculating these weights have been proposed, e.g., by using feature importance [50].

2.2.3 Latent Class Analysis

Finite Mixture Models (FMM) assume that the observed data originate from a mixture of underlying probability distributions, each corresponding to a latent subgroup [52]. The Figure 2.2 illustrates an FMM composed of three gaussian components that together form the overall distribution.

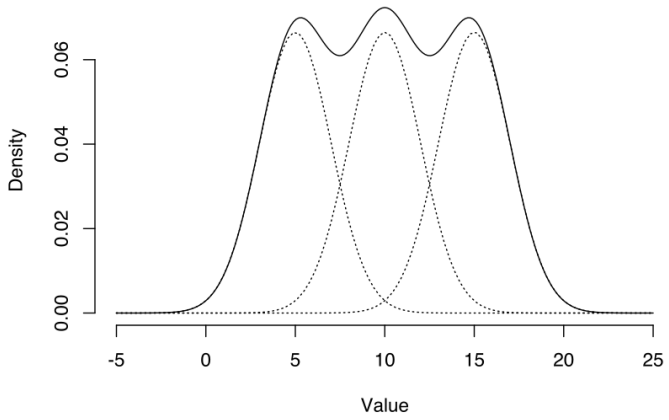

Figure 2.2: Density plot of a Gaussian Mixture Model [53]

Latent Class Analysis (LCA) is a type of FMM applied to categorical data. It uncovers latent or hidden groups based on patterns in observed covariates. For each data point, the model estimates a posterior probability for each class. The data points are then assigned to the class with the highest probability. This maximum posterior probability will be referred to as the deciding probability. In Latent Profile Analysis (LPA), all variables have a continuous distribution. A standardized nomenclature has yet to be established for the mixture of LCA and LPA, applied to mixed data [41]. Class membership probabilities in both LCA and LPA are typically estimated using maximum likelihood through the Expectation-Maximization (EM) algorithm.

In LPA, each latent class is assumed to follow a multivariate normal distribution, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iP})^\top$ is the vector of observed continuous variables for

individual i across P numerical features and K latent classes [54]. The joint mixture model is given by equation 2.4, where $\boldsymbol{\mu}_k$ is the mean vector, and a $\boldsymbol{\Sigma}_k$ is the covariance matrix of class k.

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.4}$$

Each class k has a mixing proportion π_k, that represents the probability of a random data point belonging to class k. Thus, π_k ranges from 0 to 1, $0 < \pi_k < 1$, and sum to 1 over all K classes, $\sum_{k=1}^{K} \pi_k = 1$.

On the other hand, LCA models the probability distributions of the observed variables $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iQ})^\top$ for data point i across Q categorical covariates [54]. Each latent class k defines a class-specific probability distribution over the categorical covariates. Under the assumption of local independence, the joint mixture model is given by equation 2.5.

$$P(\mathbf{u}_i) = \sum_{k=1}^{K} \pi_k \prod_{q=1}^{Q} P(u_{qi} \mid c_i = k) \tag{2.5}$$

Here, π_k is the mixing proportion for latent class k, representing the probability for data point i of belonging to that class. $P(u_{qi} \mid c_i = k)$ is the probability of a data point with class assignment c_i to cluster k responding with category u_{qi} for feature q.

While R remains the dominant language for latent class modeling, the development of the *StepMix* library, built on scikit-learn, enables the use of LCA in Python [55]. This module support a mixture of LCA and LPA, where categorical and numerical variables are modelled jointly using e.g. multinomial and normal distributions respectively, under the assumption of local independence. Even though there is no standard nomenclature for a mixture of LCA and LPA, this is simply referred to as LCA from here on.

By training LCA models with varying numbers of classes and assessing multiple validation metrics, one can choose the model with K latent classes that has the best fit, but least complexity [41]. The main fit metrics include the Bayesian Information Criterion (BIC), p-value from the Bootstrapped-Likelihood Ratio Test (BLRT), and relative entropy. All validation metrics are further described in detail in section 3.4. Proper model evaluation further requires investigation of the size of the smallest class, as very small classes might indicate overfitting. It is important that the classes represent meaningful groups, since the fit metrics alone might favor models with limited clinical relevance. In addition, model usability depends on the certainty of the group membership, making the evaluation of class probabilities important.

The best fitting model does not necessarily reveal meaningful, intrinsic patterns in the data [41]. It is therefore important to build a case for one model over another, using multiple metrics along with subject matter expertise. However, if the chosen model is not the best-fitting according to the fit metrics, it might limit replicability

on external datasets. Additionally, one should be aware of the Salsa effect, where one large scale feature dominates the classification. This can result in groups that are simply varying severities of one feature, leading to uninformative clusters. Another risk is that LCA clusters are based on the input features as the model naturally separates on these covariates. Interpretation of phenotypes based on all input features might therefore be misleading. Hence, phenotyping should be focused on the features that most clearly differentiate the groups.

2.2.4 K-Prototypes

The k-prototypes algorithm is a clustering technique designed to handle mixed datasets that consist of both numerical and categorical attributes, overcoming the limitations of traditional clustering [8]. It extends the conventional k-means algorithm, limited to numerical data, and incorporates elements from the k-modes, which facilitates clustering of categorical data.

The core idea of k-prototypes is to minimize a dissimilarity measure that accounts for both the numerical and categorical nature of the features. Similar to Gower's Distance, it uses separate distance measures for categorical and numerical data types as seen in equation 2.6 [56].

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \cdot (\sum_{p=1}^{P} |x_{ip} - \mu_{kp}| + \gamma \cdot \sum_{q=1}^{Q} \delta(x_{iq}, \mu_{kq})) \tag{2.6}$$

The cost function sums the numerical and categorical distance over n data points and K clusters, where $z_{ik} = 1$ when the data point i belongs to cluster k, and zero otherwise. The absolute distance between the data point x_{ij} and the cluster center μ_{kl} for feature l is summed over p numerical features. For categorical variables, a matching scheme δ, similar to equation 2.2 is used to compare the data point x to the cluster mode μ_{kw}, and is summed over the q categorical variables. In this algorithm [57], γ is automatically calculated for the dataset to down-weight the categorical features, and is sometimes set to the average of the numerical variables' standard deviation [58].

The k-prototypes algorithm follows an iterative procedure [59] [8]. It randomly assign K number of data points as the initial prototypes. During each iteration, each data point is assigned to the cluster whose prototype minimizes the combined dissimilarity from both numerical and categorical features. The prototypes are then updated as the mean for the numerical features and the mode for the categorical ones, forming the new centroids. The algorithm continues iterating until the cluster assignments stabilize, indicating convergence, or until a predefined maximum number of iterations is reached.

2.2.5 Validation Metrics

As an unsupervised method, clustering relies on validation to ensure usable results [8]. There is, however, no well-established best practice for cluster validation. These measures can be divided into internal metrics derived from the clustering dataset and external metrics using ground truth labels [8]. With no prior knowledge of class belonging, internal validation metrics have to be used. A common type of internal metric calculates intra-cluster similarity and inter-cluster dissimilarity, also called compactness and separation [8].

2.2.5.1 Compactness and Separation

Compactness and separation metrics can be optimized to evaluate, and choose between models, e.g, to obtain the K number of clusters [8]. The most common such metric is the Silhouette Score (SS) described in equation 2.7 as the normalized difference between the cohesion a and the separation b for the data point i.

$$SS = \frac{1}{N} \sum_{i=1}^{N} \frac{b_i - a_i}{max(b_i, a_i)} \tag{2.7}$$

Cohesion is the average distance between a data point to all other data points in its own cluster, whilst separation is the minimum average distance to the nearest cluster [21]. The SS ranges from -1 to 1, where 1 indicates that the data point is well matched to its own cluster, while -1 indicates it is closer to other clusters.

Many similar metrics present different aspects of cohesion and separation, often in weighted indexes having varying strengths and weaknesses [8]. Some are sensitive to noise, differing density, and cluster size. For average-based metrics such as the SS, clusters with an irregular shape could calculate centers which appear outside of the actual cluster, making the SS unreliable for non-convex clusters. For the SS, separation is based on the minimum distance between clusters. This can lead to a higher score when small and close clusters are merged, thus being unreliable for evaluating subclusters.

2.2.5.2 Model-fit

For model-based clustering, some validation metrics measure model-fit, such as the BIC [60]. The BIC, in Equation 2.8 is calculated from the log-likelihood $\hat{L}$ and penalized by the number of model parameters g and data points N [61].

$$BIC = g \cdot log(N) - 2 \cdot log(\hat{L}) \tag{2.8}$$

A low BIC indicates a better fit and is used to choose the best model by obtaining the number of clusters K [41].
In model-based clustering, entropy measures the certainty of the clustering [60]. It is calculated using the probability distribution P_{ik} of a variable i for cluster k as in

equation 2.9.

$$H_i = -\sum_{k=1}^{K} P_{ik} \cdot logP_{ik} \tag{2.9}$$

It takes values between 0 and $logK$, where higher values indicate greater uncertainty and more overlap between classes [62]. A caveat with this metric is that a granular model results in higher entropy simply by overfitting [41]. In model-based clustering, relative entropy is more common, where the metric is inverted and normalized, thus a higher value is preferred.

2.2.5.3 Model Testing

Another approach for determining the number of classes K in model-based methods is the BLRT that bootstraps the likelihood ratio seen in equation 2.10 [63].

$$LR = -2(logL_{K-1}(\hat{\theta_r}) - logL_K(\hat{\theta_u})) \tag{2.10}$$

Here, $\hat{\theta_r}$ and $\hat{\theta_u}$ are the maximum likelihood estimators for a simpler model with fewer parameters, and for the more complex model with more parameters, respectively. It outputs a p-value testing the LRT for K over $K-1$ classes. It has, however, been suggested that BLRT favors higher granularity to a fault, and can only be used to obtain the highest K classes to be considered [41].

2.2.5.4 The Elbow Method

As many validation metrics benefit from an increased number of clusters through overfitting, optimizing a metric by its minima or maxima can be unreliable [21]. A common method is therefore to assess the elbow, to find the point of diminishing return. The classic elbow method utilizes the Within Cluster Sum-of-Squares (WCSS) to determine the optimal number of clusters K, but can also be used with other internal metrics such as the SS. Finding the elbow is, however, a subjective method, although very common in clustering.

2.2.5.5 Cluster Tendency and Robustness

The choice of preprocessing steps, clustering model, and its parameters all influence the results, highlighting the importance of assessing clustering across multiple methods through consensus clustering [60]. Finding similar clusters across different algorithms on the same datasets additionally verifies the clustering. Another way to assess cluster tendency is through cross-validation [21]. This method splits your data into multiple parts, re-clusters it, and evaluates the cluster stability and consistency across the subsamples. Repeatability can also be evaluated by applying the clustering on a separate held-out dataset drawn from a similar population [64]. Obtaining similar clusters between different datasets means that they likely reflect an underlying structure in the data.

3
Methods

This chapter presents the methodological approach of this project, starting with the description of the dataset and the preprocessing steps applied. It further presents the clustering models that were used, the evaluation of the models, and finally, the interpretation of the results.

3.1 The Datasets and Features

This project used data from four clinical trials with a total of 11,140 patients, whose characteristics are shown in Table 3.1. The trials evaluated the effects of a drug called Dapagliflozin. DAPA-HF and DELIVER evaluated its effect on HF severity, while DETERMINE-preserved and DETERMINE-reduced assessed its effect on exercise capacity.

Trial	**Enrollment**	**Analysis**	**Subtype**
DETERMINE-r	313	241	HFrEF
DETERMINE-p	504	446	HFpEF
DELIVER	6,263	5,953	HFpEF
DAPA-HF	4,744	4,500	HFrEF

Table 3.1: Description of all trials that were used, their LVEF classification, number of patients at enrolment, and used for analysis.

All studies shared common inclusion and exclusion criteria and require NYHA class II-IV and elevated NT-ProBNP levels. Common exclusion criteria are type 1 diabetes mellitus, systolic blood pressure (SBP) ≤ 95 or > 160 mmHg, eGFR $< 25 - 30 ml/min/1.73 m^2$, recent MI, AF ablation, stroke or cardiac resynchronization therapy within the past 12 weeks, as well as heart transplantation or use of ventricular assist devices. Patients were further excluded if they had recent SGLT2 inhibitor therapy, such as Dapagliflozin, within 4-8 weeks.

Although many of the variables were the same or similar across the four trials, there were some differences. Especially as the DETERMINE trials were much smaller than DAPA-HF and DELIVER. This project used a small subset of all variables, consisting of numerical, ordinal, and binary variables from medical history data, vital signs, and baseline characteristics, see Table 3.2.

Covariate Names	Code	Type	Units/levels	Status
Age	AGE	Num	years	Clustering
Body Mass Index	BLBMI	Num	kg/m2	Clustering
NT-ProBNP	BLBNP	Num	pg/ml	Clustering
Left Ventricular Ejection Fraction	BLLVEF	Num	%	Clustering
Systolic Blood Pressure	BLSBP	Num	mmHg	Clustering
Diastolic Blood Pressure	BLDBP	Num	mmHg	Removed
Estimated Glomelural Filtration Rate	BLEGFR	Num	$ml/min \cdot 1.73m^2$	Clustering
Heart Rate	BLHR	Num	BPM	Clustering
Blood Glucose	BLHBA1C	Num	%	Removed
Sex	SEX	Bin	M/F	Clustering
New York Heart Association levels	BLNYHA	Ord	4 levels	Clustering
Myocardial Infarction	MI	Bin	Y/N	Clustering
Atrial Fibrillation	AF	Bin	Y/N	Clustering
Type 2 Diabetes Mellitus	T2DM	Bin	Y/N	Clustering
HF Diagnosis Duration	HFDIDUR	Ord	6 levels	Extra

Table 3.2: Description of all covariates that were used for initial preprocessing and clustering

Features for clustering were selected in discussion with clinical expertise at AstraZeneca. The chosen covariates represent different clinical presentations of HF, see section 2.1. Initially 14 clinical covariates were chosen, of which 9 numerical, 4 binary, and 1 ordinal.

In addition to the main clustering model, for further analysis, heart failure duration of diagnosis (HFDIDUR) was added as an ordinal covariate to evaluate its effect on clustering. In addition to the clustering covariates, all-cause mortality-, cardiovascular (CV) death- and a primary endpoint were used for survival analysis using Kaplan-Meier curves. The primary endpoint was constructed as a composite of CV death, CV hospitalization, and CV events requiring a hospital visit. Some variables, such as CAD, LDL cholesterol, and KCCQ were initially considered for clustering, but were excluded as they were unavailable in some datasets. For KCCQ, NYHA was chosen as a substitute.

3.2 Preprocessing

After selecting the covariates, the preprocessing began by extracting baseline entries and outcome data for randomized patients. Preliminary assessment of the data revealed missing values and duplicate records. Duplicates were identified and eliminated. Missing data for the chosen covariates accounted for only 0.1% of the

data. Although imputation was possible, and model based methods can handle missing data, a complete-case approach was used as the amount of missing data was negligible, thereby avoiding any potential bias associated with imputation methods.

Following the creation of the complete case dataset, ordinal variables with categories representing less than ten percent of samples were collapsed with adjacent levels to minimize potential bias, due to sparse distributions. For BLNYHA, categories I and II were merged, as were III and IV, due to negligible frequencies in classes I and IV. The higher severity group (III/IV) was encoded as 1 and the lower severity group (I/II) as 0, thereby preserving their ordinal nature. This process was repeated for HFDIDUR where 7 ordinal categories were compiled into 6. Further, binary categorical variables such as AF, MI, T2DM, SEX, and BLNYHA, were numerically encoded for compatibility with the clustering methods and validation metrics that were used.

After data cleaning and encoding, covariates were visually inspected using histograms and density plots to evaluate normalcy. Numerical covariates, such as BLBNP, BLHBA1C, BLBMI, BLHR, BLDBP, and BLSBP, which exhibited right-skewed distributions, were log-transformed as they deviated from the assumption of symmetry required by the clustering methods. To further evaluate the assumption of normality, a statistical test such as Shapiro-Wilk was employed, however because of the size of the dataset such statistical tests were of limited use.

To further reduce skewness, preprocessing of outliers and feature scales were considered. One variable, BLBNP, contained extreme values exceding 40,000 whilst the mean was 1,877 and the standard deviation 2,475. These extreme values could have been clipped during preprocessing. However, they were clinically feasible for patients with severe heart failure and excluding them could have resulted in loss of relevant patient information. Similarly for other features, all outliers were clinically feasible and no values were excluded. To handle differences in feature scales, normalization was employed. All continuous variables were standardized by Z-score to have a zero mean and unit variance, thus preventing any single attribute from dominating clustering.

To ensure the independence of input features, correlation analysis using Pearsons's coefficient was performed. In cases of correlation > 0.5, one variable from each pair was removed. When both variables in a collinear pair were equally important, sensitivity analyses were conducted. Each variable was excluded and the models re-run, where improvement in model performance resulted in retaining the more informative variable.

3.3 Clustering

This section presents the clustering algorithms considered for analysis. K-medoids, agglomerative hierarchical, LCA and k-prototypes were all implemented and assessed for eligibility in use with mixed data. As k-medoids and hierarchical used

Gower's distance to enable mixed data, the categorical features were dominating. Therefore LCA was chosen as the main model and k-prototypes for sensitivity analysis.

3.3.1 Latent Class Analysis

LCA was implemented on the standardized clinical data using the *StepMix* library in Python. To determine the number of latent classes that best represent the underlying data structure, the models were trained with 2 to 10 clusters. The seven numerical features were modeled using independent gaussian distributions. In *StepMix*, the *'continuous'* measurement model estimated the mean and variance of each continuous variable in each class. In contrast, the five categorical features were modeled using a *'bernoulli'* measurement model. When HFDIDUR was used, the *'categorical'* option was chosen to estimate a multinoulli distribution.

The settings chosen for LCA allowed for fuzzy class membership (soft clustering) meaning individuals retained their probabilistic membership across classes. Further, the number of initializations were set to 50, and the best fitting solution was kept. Additionally, the maximum number of EM iterations was set to 100,000 per initialization, improving convergence and avoiding local solutions. In the fitted model, the posterior class probabilities were calculated using *predict_proba_class*, while hard cluster labels were generated using *predict* which assigns each patient to the class with the highest probability.

To evaluate the contribution of LVEF, an ablation analysis was performed by removing it from the data and re-evaluating the models. This was done to estimate whether similar latent classes were obtained in the absence of LVEF, and if clinically recognized classes such as as HFrEF, HFmrEF and HFpEF naturally occur without it. HFDIDUR was also added for re-clustering to examine wether the clusters exhibited consistent patterns when stratified by the duration of heart failure.

3.3.2 K-Prototypes

K-prototypes was used as a sensitivity analysis to validate the models found in LCA. Clustering was performed over a range of 2 to 10 clusters to identify the best fitting model, and compare it to the main LCA results. The *KPrototypes* function from the *kmodes* Python package was used [57]. To differentiate the numerical from the categorical covariates, the column indices of the categorical features were explicitly given as input to the model. For each k-cluster model, the algorithm was fitted and hard cluster assignments were calculated using *predict*. Additionally, WCSS was calculated as a measure of cluster compactness, later used for analysis via the elbow method.

3.4 Validation

After training the models and assigning data points to clusters, the resulting groupings were first visualized using PCA, UMAP, and 3-by-3 covariate plots. The 3-by-3 plots displayed scatter plots for every combination of three covariates, thus giving a visual impression of whether any three covariates separates the clusters. The first validation metric to be calculated was the silhouette score, using a function from scikit-learn. As the dataset contains mixed data types and the SS requires a distance metric, a precomputed Gower's distance matrix from the *gower* package was used as input. As the SS relies on comparing data points between its own and other clusters, it could only be calculated for $k > 1$, and all models were therefore only trained to fit that constraint.

The BIC, relative entropy and BLRT-sweep functions were employed from the StepMix packages to evaluate LCA. The SS, BIC's and relative entropies were thereafter plotted for LCA models with 2 to 10 clusters. The validation metric curves were then evaluated for the highest SS and entropy along with the lowest BIC, which were evaluated to choose the number of clusters. The elbow method was used where the optimal value was uninformative, e.g. when the BIC is decreasing as the number of clusters increase. Furthermore, the p-value for the BLRT was assessed for the maximum number of clusters to consider. The highest deciding probabilities were also plotted for each cluster where a larger proportion of high probability assignments means a certain clustering, which can inform the choice of K. All models were also further evaluated based on the sizes of the clusters, where very small clusters should be avoided.

For the models that were chosen, the cluster descriptions were displayed. Histograms were plotted for each cluster and covariate, thus visualizing the differences in cluster characteristics. Further, tabular descriptions were constructed where each numerical covariates were described by mean and standard deviation, and the categorical variables by frequencies. These tables were used to observe any covariates that were distinct for each cluster. From this, the initial cluster descriptions were obtained, and through discussion with clinical expertise, refined to ensure that the descriptions were clinically feasible. For alternative visualization, a covariate heatmap was also created from this tabular data. This was done through dichotomizing numerical variables by levels that are relevant for diagnosis. One example is that values of BMI above $30 kg/m^2$ are defined as obese, therefore, the BLBMI variable was dichotomized by the clinical cutoff of BMI > 30. This resulted in a dark coloration in the heatmap if more patients in that cluster had clinically significant values.

For these chosen clusters, the cardiovascular death, all-cause mortality, and primary composite endpoint were used to plot a Kaplan-Meier survival curves, to assess if the clusters differed in outcome. These curves plot the survival function to evaluate how the risk of the event changes over time. Input from cardiologists were used to evaluate clinical relevance and feasibility. The clusters were then also compared to those found in literature.

4

Results

This chapter presents the results from the clustering analyses, including evaluation of model performance using internal validation metrics, selection of the optimal number of clusters, description of the resulting clusters, their clinical interpretation, and outcome.

4.1 Exploratory Data Analysis

In Figure 4.1, the distribution of the dataset is shown before preprocessing. Histograms or bar charts are plotted for each covariate, subdivided for each dataset. For the HFrEF trials (DETERMINE-r and DAPA-HF), LVEF curves are distinctly lower than the HFpEF trials (DETERMINE-p and DELIVER). The bar chart for SEX shows that the dataset has more males than females for every study, specifically 64.6% males in the entire dataset.

Figure 4.1: Histogram distribution per dataset before clustering

4.1.1 Preprocessing

Figure 4.2 shows univariate distributions of all continuous covariates before and after preprocessing. Before preprocessing, skewed distributions of BLBNP, BLHBA1C, BLBMI, BLHR, BLDBP, and BLSBP can be observed. The complete dataset is cleaned and standardized dataset with a log-transform applied to the aforementioned covariates, for which, HBA1C is still notably skewed.

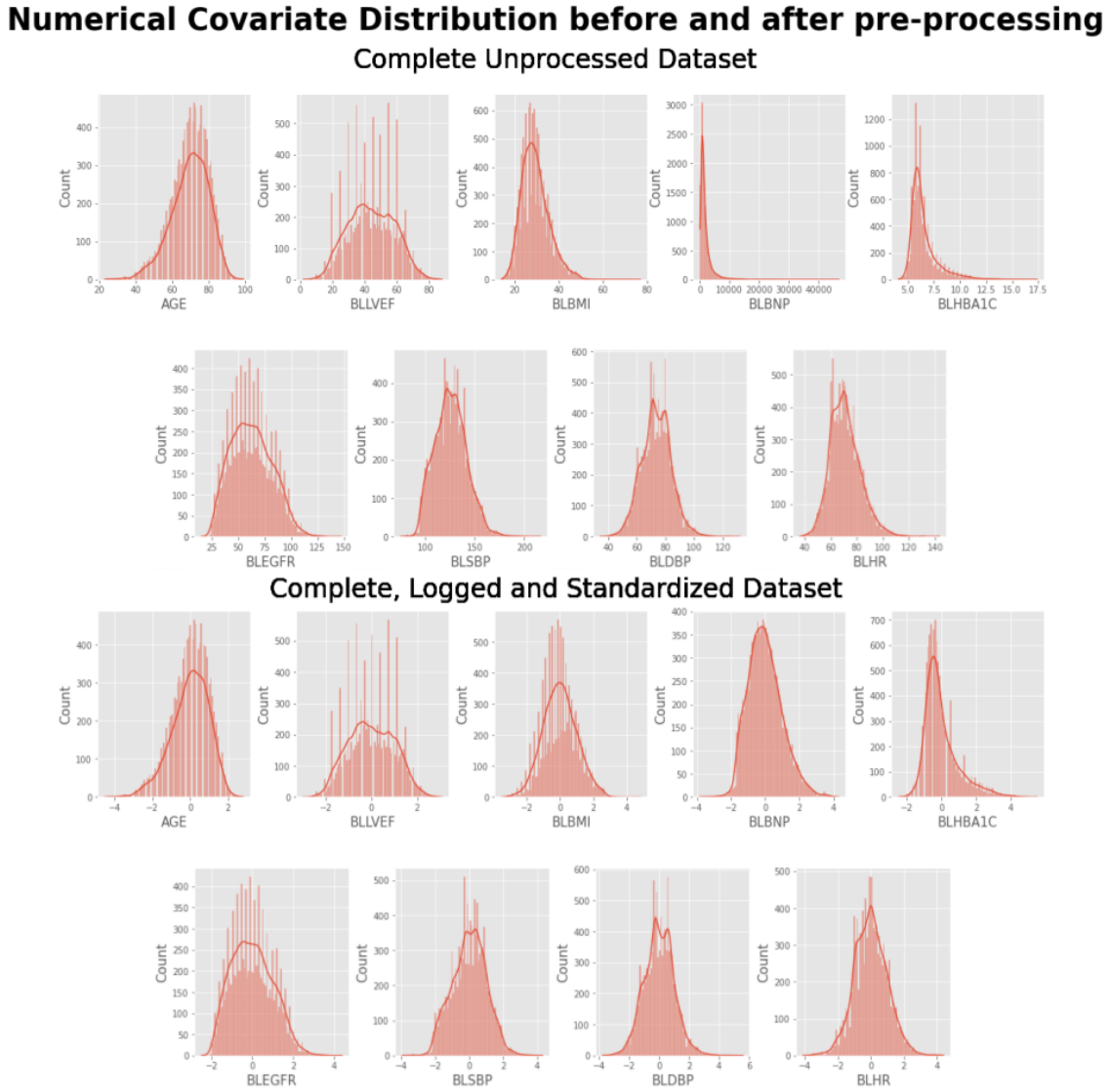

Figure 4.2: Distribution of continuous covariates before and after pre-processing

Normality was further tested using the Shapiro-Wilk test, which all covariates failed, indicating non-normality. However, the transformed covariates display sufficiently bell-shaped profiles that satisfy the practical assumptions for mixture modeling.

4.1.1.1 Correlation

In Figure 4.3, Pearson's correlation matrix is shown for all clustering covariates, including HFDIDUR. Two sets of features showed a correlation > 0.5. T2DM and BLHBA1C had a correlation of 0.57, for which BLHBA1C was removed as it had a non-normal distribution after preprocessing. BLSBP and BLDBP had a correlation of 0.53, for which BLDBP was removed. It was also observed that AGE had a high correlation of -0.43 with BLEGFR.

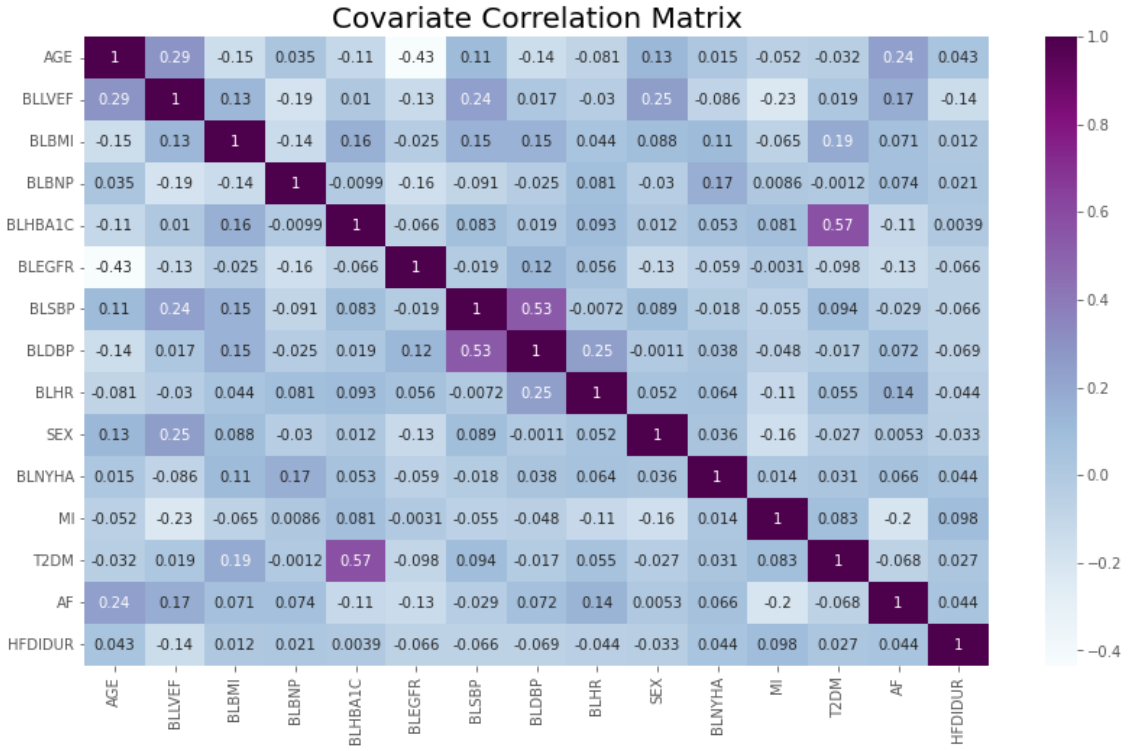

Figure 4.3: Correlation matrix

4.2 Clustering with Latent Class Analysis

This section presents the main result of this project, along with clustering without LVEF for further comparison. For extended and additional model results, refer to Appendix A.

In Figure 4.4, the BIC scores, relative entropies, and silhouette scores are presented for the 2- to 10-cluster models. The BIC score continuously decreases, and an elbow can be observed at the 4-cluster model for a BIC of 282,767. The maximum relative entropy value is observed at the 5-cluster model for a value of 0.669, after which the score decreases. The silhouette score has its highest value of 0.128 at the 2-cluster model , but the peak to a SS of 0.085 at 4-clusters, indicates this model as the better choice.

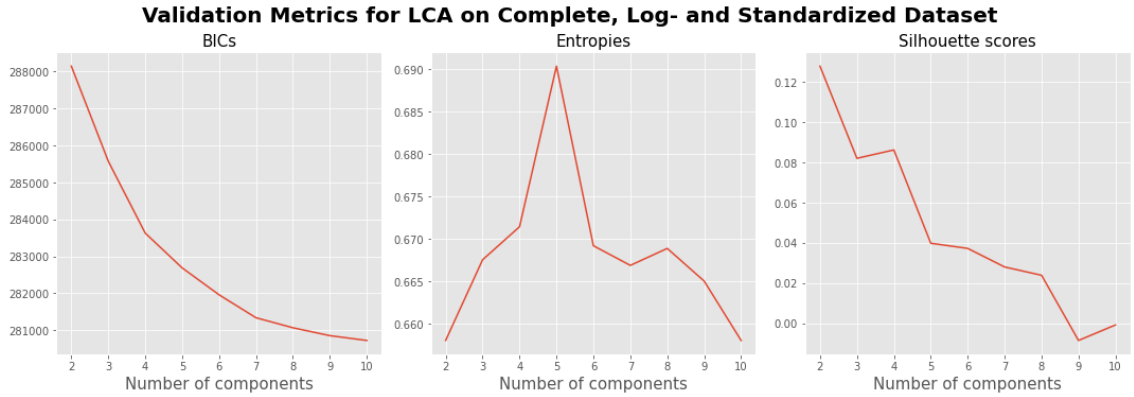

Figure 4.4: Validation Metrics for LCA

While performing the BLRT sweep and testing $K-1$ and K classes, the first significant p-value of 0.16 was obtained at 9 vs 10 classes, indicating that the maximum number of classes to consider is 9. Because of the BIC elbow at 4 clusters and SS peak for that same model after the first elbow, the 4-cluster model is the model of most interest.

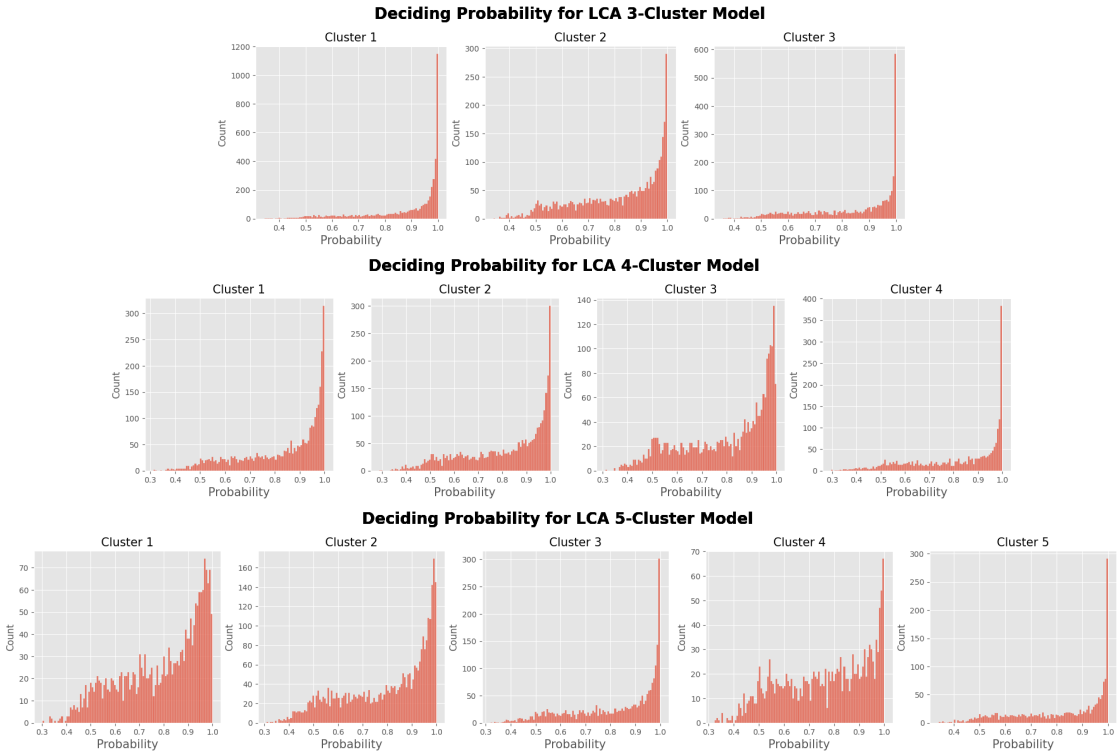

Figure 4.5: The deciding probabilities for LCA 3, 4, and 5-cluster models.

In Figure 4.5, the deciding probabilities for the LCA 3-, 4-, and 5-cluster model can be observed. For the 3-cluster model, although all clusters are left-skewed, clusters 1 and 3 seems the most certain. In the 4-cluster model, cluster 3 appears to be the most uncertain, while clusters 1 and 2 have partially ambivalent maximum posterior

probabilities. As the number of clusters increases, the deciding probabilities decrease indicating less certain class assignments. For the 5-cluster model, the probabilities look significantly less left-skewed, especially for clusters 1 and 4, further validating the choice of the 4-cluster model.

The distributions for each covariate are presented as histograms for each cluster, in Figure 4.6. Here, AGE, BLLVEF, BLBNP, BLEGFR, AF, MI, and SEX are visually distinguishing the four clusters. The distribution per covariate is further presented in Table 4.1, for the LCA 4-cluster model. In this visualization of the four clusters, one can observe distinct separation on AGE, SEX, BLEGFR, BLBNP, MI, AF, and BLLVEF. **Cluster 1** (29.1%) is the older cluster with the most females, no prevalence of myocardial infarction, highly prevalent atrial fibrillation, and preserved ejection fraction. **Cluster 2** (29.2%) is an old male cluster with the lowest BMI, lowest eGFR, highest NT-ProBNP, highest prevalence of myocardial infarction, reduced ejection fraction, and the longest heart failure diagnosis duration. **Cluster 3** (22.0%) is a cluster with high BMI, low eGFR, low NT-ProBNP, the highest prevalence of diabetes, the lowest prevalence of atrial fibrillation, and a preserved ejection fraction. **Cluster 4** (19.8%) consists of younger males, with high eGFR, normal blood pressure, and reduced ejection fraction.

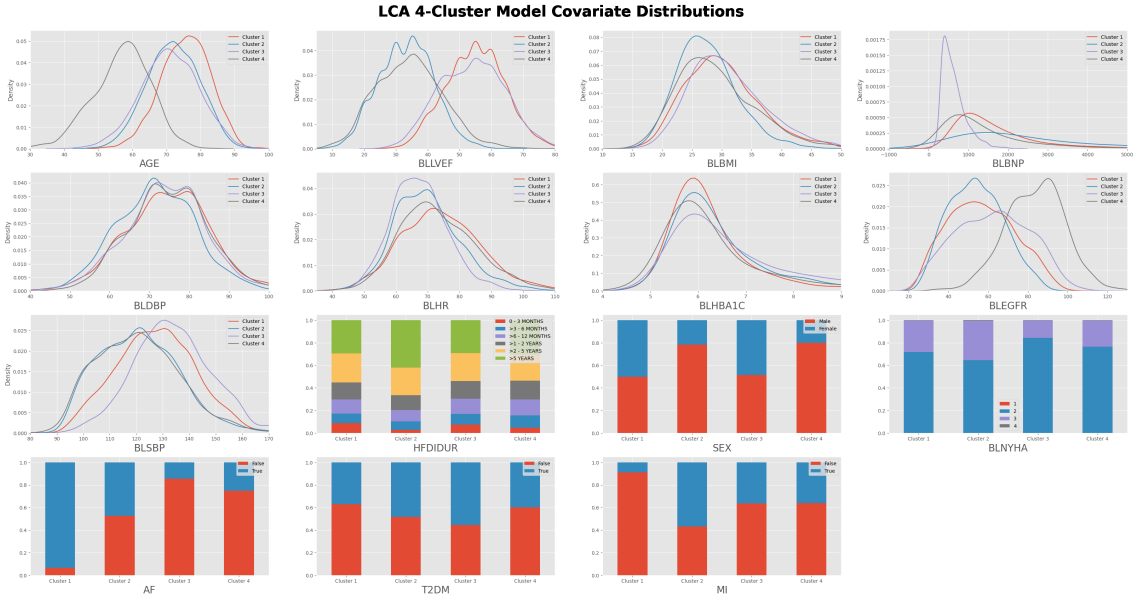

Figure 4.6: Covariate histograms per cluster for the LCA 4 cluster model

Figure 4.7 shows the Kaplan-Meier survival analysis curves for 4 latent classes on cardiovascular death, all-cause mortality, and primary composite endpoint. When observing the survival curve for cluster 2, they have the highest risk of event for the primary endpoint, CV-death, and all-cause mortality. On the contrary, cluster 3 has the lowest risk of event across the primary endpoint, CV-death, and all-cause mortality. Notably, the clusters with HFrEF (clusters 2 and 4) have a higher CV-death and primary endpoint than HFpEF. When assessing all-cause mortality, clusters 1, 3, and 4 have a very similar risk of event.

Covariate	Cluster 1	Cluster 2	Cluster 3	Cluster 4
AGE	76.0 ± 7.19	72.0 ± 7.58	71.0 ± 8.48	57.0 ± 8.58
BLLVEF	55.0 ± 9.14	34.0 ± 8.52	55.0 ± 10.21	35.0 ± 10.35
BLBMI	29.4 ± 6.23	27.0 ± 5.09	30.0 ± 6.23	28.0 ± 6.68
BLBNP	1362.0 ± 1247.56	2017.18 ± 3809.21	557.0 ± 301.39	1025.05 ± 1629.91
BLHR	73.5 ± 12.97	69.0 ± 10.81	67.0 ± 8.9	71.67 ± 12.08
BLEGFR	56.0 ± 16.42	54.0 ± 13.62	63.0 ± 18.42	84.0 ± 14.95
BLSBP	127.5 ± 14.77	120.67 ± 15.55	133.0 ± 14.83	120.0 ± 15.9
BLHBA1C	6.1 ± 1.08	6.2 ± 1.27	6.3 ± 1.57	6.0 ± 1.59
BLDBP	75.0 ± 10.62	71.67 ± 10.11	74.33 ± 10.16	74.67 ± 10.34
HFDIDUR (>2 Years)	55.04%	66.57%	54.05%	53.66%
SEX (Female)	50.06%	21.49%	48.48%	19.87%
BLNYHA (III/IV)	28.03%	35.39%	15.6%	23.33%
AF (True)	93.66%	47.39%	14.5%	25.24%
T2DM (True)	37.07%	48.46%	55.53%	39.84%
MI (True)	8.88%	56.92%	36.28%	36.15%

Table 4.1: Cluster covariate table for main LCA 4-cluster model - Numerical covariates are presented as mean ± standard deviation; categorical covariates are shown as percentages based on observed fractions. HFDIDUR, BLDBP, and BLHBA1C were not used in clustering.

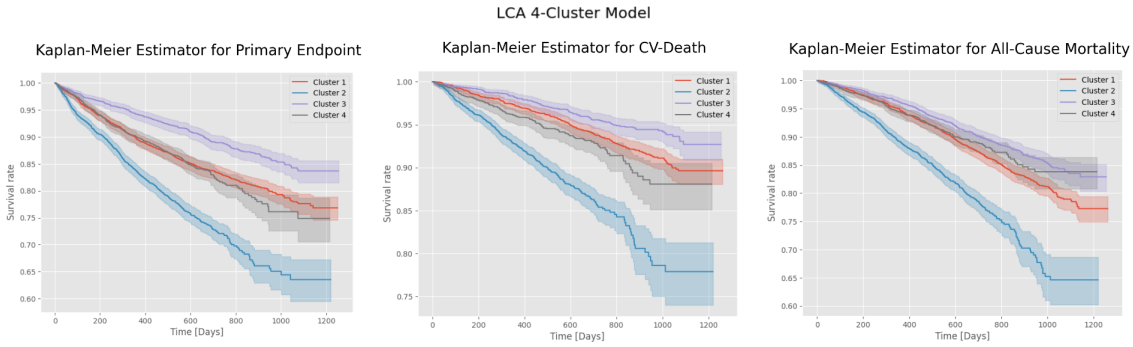

Figure 4.7: Kaplan-Meier estimator for LCA 4-Cluster model outcomes. Primary endpoint indicates CV - death, hospitalization, and events that require a hospital visit.

4.3 Clustering without LVEF

This section evaluates if the standard LVEF subtypes of HFpEF, HFmrEF and HFrEF, are naturally occuring, when clustered without LVEF. In Figure 4.8 the validation metrics for the LCA model without LVEF are presented. The BIC has a slight elbow at 255,164 for the 4-cluster model. The entropies increase up until the 5th cluster with a relative entropy of 0.629, after which it slowly decreases. The SS

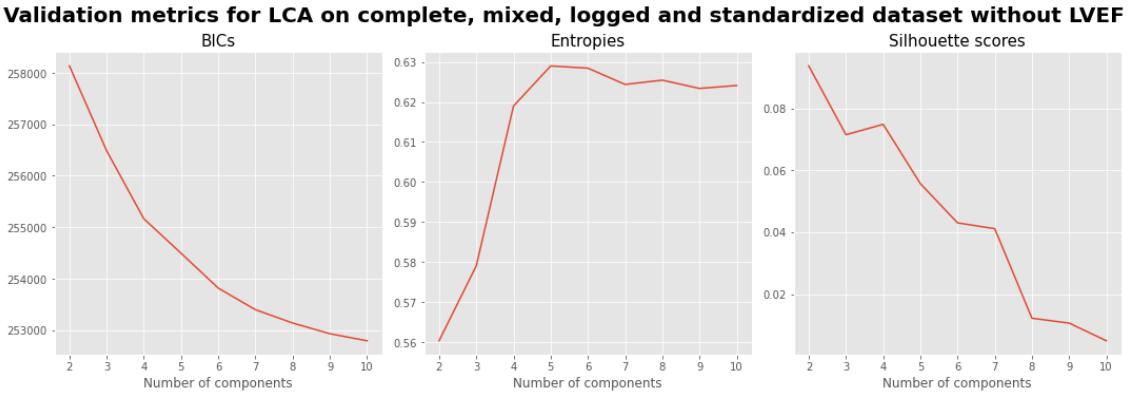

Figure 4.8: Validation Metrics for LCA without LVEF

has its elbow at 0.071 for the 3-cluster model and a slight increase to 0.075 for the 4-cluster model. The 4-cluster model is of most interest here.

In Figure 4.9, the deciding probabilities can be observed for the 3-, 4-, and 5-cluster LCA model without LVEF. These results are less left-skewed than for the main LCA models. For the 3-cluster model, while clusters 1 and 3 are left-skewed, cluster 2 is significantly less so. This indicates less certainty if compared to Figure 4.5. Similar observations can be made of the 4- and 5-clusters.

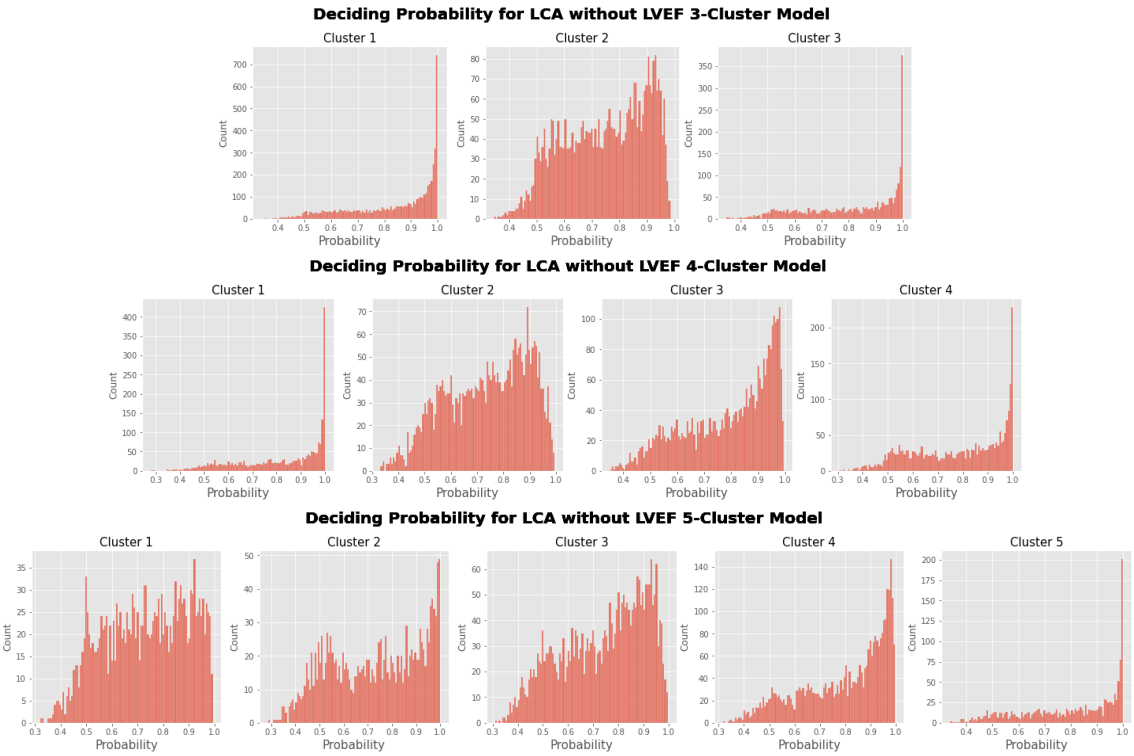

Figure 4.9: The deciding probabilities for LCA model without LVEF for 3, 4 and 5-clusters.

In Figure 4.10, the LVEF histograms for the 2-, 3-, and 4-cluster models are shown. The 2-cluster model subdivides LVEF into two groups, specifically LVEF with mean 40%, corresponding to HFrEF, and 47% corresponding to HFmrEF.

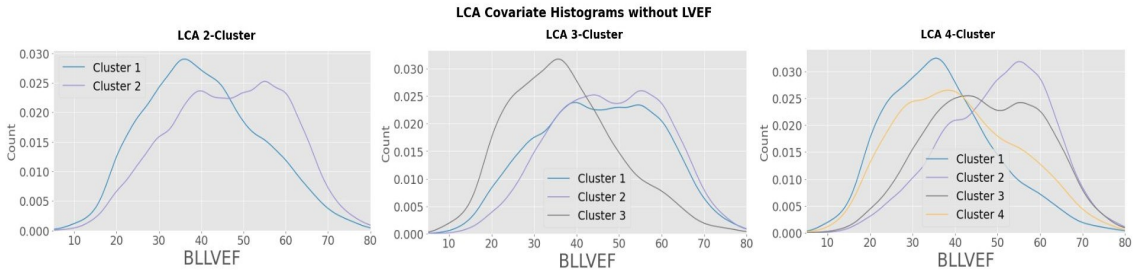

Figure 4.10: LVEF distribution histograms for 3-, 4-, and 5 cluster LCA models

The 3-cluster model creates two very similar groups based on LVEF with a mean of 45% and 48%, thus characterizing them as HFmrEF, and one distinct HFrEF group with a mean of 37%. The 4-cluster model subdivides into a lower LVEF with a mean of 37% corresponding to HFrEF, and 3 groups with higher LVEF with means of 41, 47 and 50%.

4.4 Sensitivity Analysis using K-Prototypes

For validating the four clusters found in section 4.2, clustering was performed on the same data and covariates using the k-prototypes algorithm.

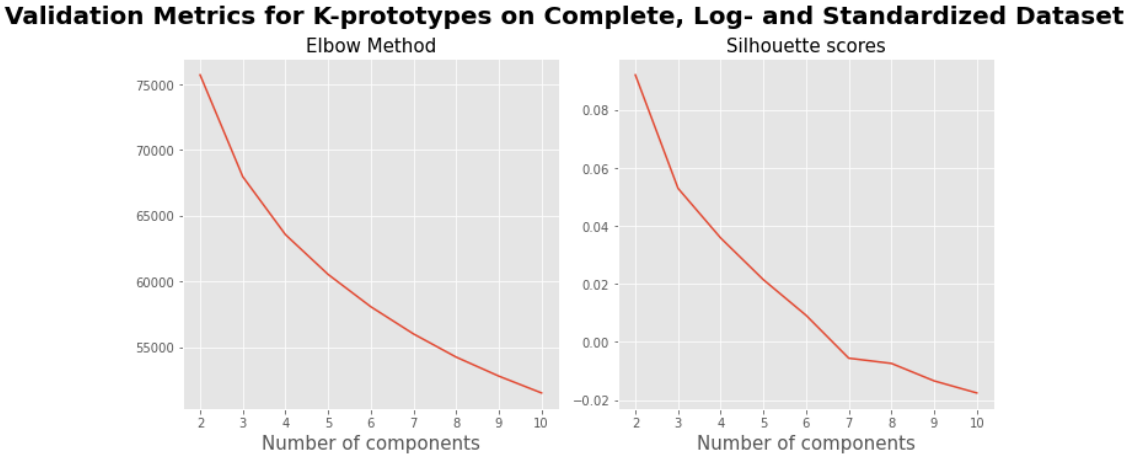

Figure 4.11: The validation metrics for k-prototypes, including the silhouette score and elbow method (WCSS)

The validation metrics for k-prototypes, including the inertia and Silhouette score, are shown in Figure 4.11. The inertia plot has the most distinct elbow at WCSS = 67,967 for the 3-cluster model. The silhouette score, which has a maximum value of 0.092 at the 2-cluster model, has an elbow at the 3-cluster model for a SS of 0.053. Therefore, the 3-cluster model is of interest according to the validation

metrics, however the 4-cluster model is also presented for comparison to the main LCA model.

4.4.1 3-Cluster Model

In this section, the result for the 3-cluster k-prototype model will be presented. Firstly, in Figure 4.12, the distribution of the covariates is presented, stratified by the clusters.

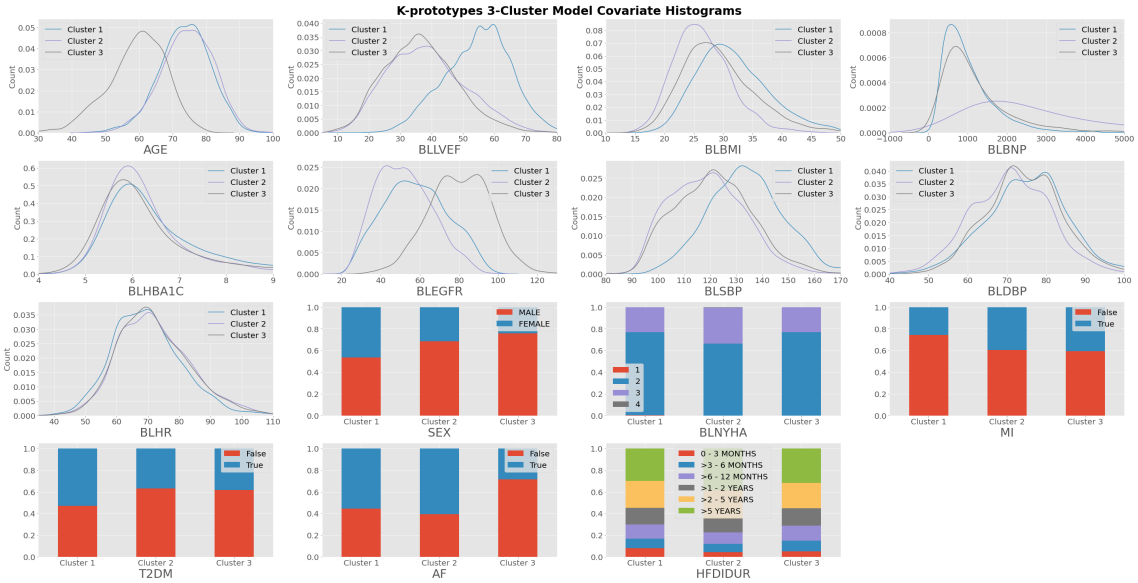

Figure 4.12: The covariate histograms per cluster for the 3-cluster k-Prototype model

There is a distinct difference between the clusters when assessing based on BLLVEF, AGE, BLEGFR, and AF. There is a slight distinction between clusters for BLBMI, BLSBP, BLBNP, SEX, BLNYHA, MI, T2DM, and HFDIDUR, while BLDBP, BLHBA1C, and BLHR give no visual distinction. Further, in Table 4.2, the dichotomized covariates are shown for each cluster for the k-prototypes 3-cluster model. **Cluster 1** (39.8%) is the largest in the cohort, with approximately 40% of the patient population, with an older, mixed-gendered profile, exhibiting the highest BMI, a high proportion of diabetes, the lowest proportion of AF and MI, the lowest NT-ProBNP, and preserved ejection fraction. **Cluster 2** (31.5%) is the oldest cluster with the highest symptom burden, the lowest BMI, the lowest eGFR with impaired renal function, the highest NT-ProBNP, the highest prevalence of diabetes, and reduced ejection fraction. **Cluster 3** (28.8%) comprises the youngest individuals, mostly male, with a low proportion of AF, but high MI, the highest eGFR, and reduced ejection fraction.

Covariate	Cluster 1	Cluster 2	Cluster 3
AGE	73.61 ± 7.55	74.36 ± 7.73	59.37 ± 8.72
BLLVEF	54.67 ± 10.46	38.33 ± 12.24	36.52 ± 11.3
BLBMI	31.65 ± 6.06	26.43 ± 4.79	29.22 ± 6.22
BLBNP	1031.81 ± 710.99	3476.27 ± 3703.3	1299.66 ± 1223.21
BLHBA1C	6.67 ± 1.38	6.38 ± 1.15	6.54 ± 1.56
BLEGFR	58.03 ± 16.22	52.12 ± 14.14	80.18 ± 15.63
BLSBP	133.87 ± 14.6	119.1 ± 14.33	121.57 ± 14.85
BLDBP	75.17 ± 10.49	71.13 ± 10.21	74.57 ± 9.93
BLHR	69.83 ± 11.24	72.49 ± 12.17	72.05 ± 11.57
SEX (Female)	46.39%	31.69%	24.27%
BLNYHA (III/IV)	23.29%	33.57%	23.3%
MI (True)	25.59%	39.63%	40.6%
T2DM (True)	53.02%	36.89%	38.23%
AF (True)	55.37%	60.51%	28.42%
HFDIDUR (>2 Years)	54.67%	64.63%	55.15%

Table 4.2: Cluster covariate table for k-prototypes 3-cluster model - Numerical covariates are presented as mean ± standard deviation; categorical covariates are shown as percentages based on observed fractions

4.4.2 4-Cluster Model

The k-prototypes 4-cluster model is presented in this section for comparison with the main result.

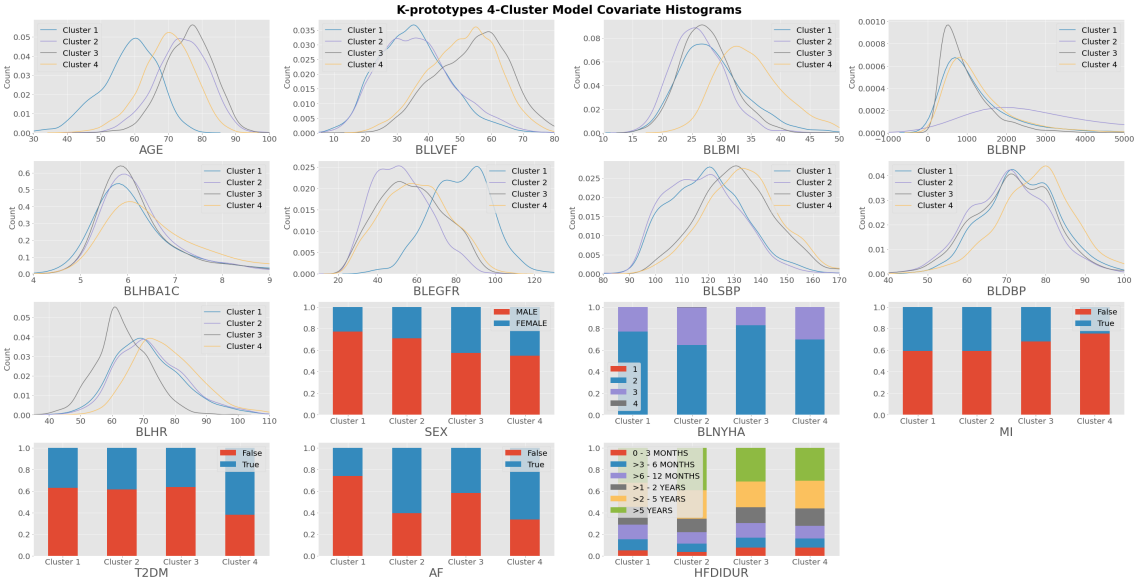

Figure 4.13: Covariate histograms per cluster for the k-prototypes 4-cluster model

In Figure 4.13, the distributions for each covariate are presented as histograms for each cluster. Here, AGE, BLLVEF, BLBNP, BLEGFR, AF, BLSBP, and SEX vi-

sually differentiate the four clusters.

Covariate	Cluster 1	Cluster 2	Cluster 3	Cluster 4
AGE	58.19 ± 8.63	73.98 ± 7.81	76.02 ± 7.16	69.77 ± 7.78
BLLVEF	35.14 ± 11.08	36.29 ± 11.76	53.91 ± 11.56	50.85 ± 10.89
BLBMI	28.52 ± 5.76	26.13 ± 4.45	27.5 ± 4.45	34.88 ± 5.68
BLBNP	1311.3 ± 1219.11	3864.46 ± 3978.76	967.79 ± 682.53	1335.93 ± 1056.07
BLHBA1C	6.49 ± 1.52	6.4 ± 1.17	6.33 ± 1.13	6.92 ± 1.56
BLEGFR	82.09 ± 15.44	52.34 ± 14.28	57.57 ± 16.35	59.99 ± 16.61
BLSBP	119.66 ± 14.4	117.94 ± 13.98	131.01 ± 14.73	133.37 ± 15.04
BLDBP	73.98 ± 9.73	70.86 ± 10.03	72.05 ± 10.02	77.93 ± 10.32
BLHR	71.54 ± 11.13	72.7 ± 11.42	63.46 ± 8.37	77.41 ± 10.98
SEX (Female)	22.97%	29.23%	42.67%	45.33%
BLNYHA (III/IV)	22.89%	35.33%	17.17%	30.32%
MI (True)	40.77%	40.82%	31.9%	24.66%
T2DM (True)	37.08%	38.38%	36.28%	61.95%
AF (True)	26.08%	60.42%	41.75%	66.2%
HFDIDUR (>2 Years)	54.76%	65.63%	55.17%	55.98%

Table 4.3: Cluster covariate table for k-prototypes 4-cluster model - Numerical covariates are presented as mean ± standard deviation; categorical covariates are shown as percentages based on observed fractions

Table 4.3 shows the covariate descriptions for the k-prototypes 4-cluster model. **Cluster 1** (23.1%) consists of a population with reduced ejection fraction, characterized by a younger demographic, mostly male, with better kidney function, with the least prevalence of AF, and shortest duration of HF. **Cluster 2** (25.4%) is an elderly, predominantly male group with the highest NT-ProBNP, lowest BMI, severely impaired renal function, highest symptom burden, longest duration of HF, and reduced ejection fraction. **Cluster 3** (25.4%) is the oldest group with the lowest NT-ProBNP, mixed gender, lowest heart rate, prevalence of diabetes, and symptom burden, along with a preserved ejection fraction. **Cluster 4** (26.1%) shows a mix of male and female patients, with the highest BMI, blood pressure, and heart rate, but the lowest proportion of MI, highest proportion of AF and T2DM, along with preserved ejection fraction.

4.5 Discussion

Latent class analysis was chosen as the main clustering algorithm for this project as it inherently handles missing and mixed data, without categorical covariates dominating or misconstruing similarities. It further models latent distributions and assigns fuzzy class memberships, which are useful characteristics for patient data, as phenogroups might have overlapping distributions. This work confirmed LCA to be effective in clustering for this type of data through its validation metrics being better than for other methods.

From the main LCA model, **Cluster 1** has the oldest patients with the highest prevalence of AF, thus driving up the NT-ProBNP. They also have a very low frequency of prior MI. As discussed in Section 2.1, HFpEF is characterized by multiple comorbidities such as hypertension and CKD, which is also evident in this cluster from the elevated systolic blood pressure and low eGFR. **Cluster 2** represents an old and predominantly male cohort. The extremely high NT-ProBNP levels go hand in hand with their prevalence of MI. This, in combination with their lower BMI and longer HF disease duration, suggests a very ill and old group for which some patients have become cachexic. This is reflected in their high mortality compared to the other groups. Further, the low eGFR could be a result of a bidirectional comorbid relationship between CKD and systemic cardiovascular disease. **Cluster 3** is an old patient group characterized by HFpEF, with the highest prevalence of obesity and T2DM. This could be interpreted as a cardiometabolic phenotype, possibly developing heart failure through diabetic cardiomyopathy. The NT-ProBNP levels of these patients could be artificially lowered, which has been observed in obese patients in previous research. **Cluster 4** represents a well recognized phenotype, with the youngest patients, early-onset HFrEF and a more severe type of HF. They are distinguished by normal blood pressure and relatively high eGFR, owing to a lower age and shorter duration of heart failure. Subsequently, they have a higher CV-mortality, although being the youngest. Due to their young age, all-cause mortality is the lowest among all clusters.

Based on the survival curves, these results do seem to make clinical sense. Especially for the HFrEF groups, where Cluster 2 has the worst outcomes, probably as they are the oldest HFrEF group, allowing for comorbidities accumulate and worsen over time. Similarly, Cluster 4 has very high incidence of primary endpoint and CV-death for being the youngest cohort. However, all-cause mortality is similar for clusters 1, 3, and 4.

To set the observed clusters into perspective, the clusters identified across the full LVEF spectrum in recent literature are presented in Table 4.4. Of the papers mentioned, Jasinska-Piadlo et. al. [20] found 4 clusters, Urban et. al. [13] found 6 clusters, and Gu et.al [17] found 3 clusters. Urban et. al found a younger cluster partially corresponding to cluster 4 in this work. This is a common group shown in literature, and the most certain and distinct in this project. They also found a cluster with advanced age, hypertensive, diabetic, low renal function and recent HF

that shared traits from cluster 3. An additional cardiorenal HFrEF cluster presented in their paper matched with cluster 2.

Source and Parameters	**Method and Data**	**Cluster Characteristics**
This thesis, N=11,140, K=4, F=12	Mixed data, LCA, DELIVER, DETERMINE and DAPA-HF trials	**Cluster 1**: Old AF no MI (HFpEF). **Cluster 2**: Male high NT-ProBNP with MI (HFrEF). **Cluster 3**: Obese diabetic (HFpEF). **Cluster 4**: Young male, with high eGFR (HFrEF).
Urban et. al 2022 [13], N=381, K=6, F=63	K-medoids with numerically encoded mixed data from acute HF registries	**Cluster 1:** Most new HF, with preserved renal function. **Cluster 2:** Old, hypertensive, diabetic, advanced atherosclerosis and comorbidities, low renal function, significant new HF **Cluster 3:** Young "healthy", early-stage HF, presumed toxic etiology **Cluster 4**: HFrEF with reduced iron resources. **Cluster 5:** Men, HFrEF, with cardiorenal syndrome, hyperventilation, right-ventricular failure. **Cluster 6:** HFpEF with increased inflammatory markers.
Gu et. al 2025 [17], N=343, K=3, F=29	K-means, hierarchical, GMM from EHR and digital twin data	**Cluster 1:** High vessel compliance and/or low vascular congestion **Cluster 2:** High pulmonary vascular resistance or pericardial constraint **Cluster 3:** Severe right- and/or left-ventricular systolic dysfunction
Jasinska-Piadlo et. al 2023 [20], N=635, K=4, F=68	Numerical covariates on k-means from EHR data (Data-driven)	**Cluster 1:** Female, low comorbidity, NT-ProBNP, HFpEF, good renal function. **Cluster 2:** Mostly males, HFrEF, highest NT-ProBNP, prevalent MI, lung and liver disease. **Cluster 3:** Female, oldest, HFpEF, lowest NT-ProBNP, prevalent peripheral vascular disease, CKD, solid tumor, and peptic ulcer. **Cluster 4:** Female, youngest, HFpEF, low NT-ProBNP, highly prevalent diabetes, and dementia.
Jasinska-Piadlo et. al 2023 [20], N=635, K=4, F=7	Numerical covariates on k-means from EHR data (Domain-driven)	**Cluster 1:** High NT-proBNP, HFrEF, less CKD, severe symptoms. **Cluster 2:** Mostly female, lowest NT-ProBNP, HFpEF. **Cluster 3:** Mostly male, old, highest NT-ProBNP, HFmrEF, most MI, lung-, and liver disease, and dementia. **Cluster 4:** Mostly female, old, low NT-ProBNP, HFpEF, peripheral valvular disease, stroke, CKD, and diabetes.

Table 4.4: Table for comparison of methods and results in this project with that of recent literature that phenotypes on the full LVEF spectrum. N-number of data points, k-number of clusters, F-number of features.

It is important to observe how well separated, homogeneous, repeatable, and usable these clusters are. As the obtained silhouette score is 0.072 for the chosen model, this indicates that it is not separating the data well. Further, the silhouette score indicates higher cohesion for fewer clusters. This is further shown by a BIC score that is ever decreasing, possibly indicating that there is no best solution other than increasing the model granularity to overfit to the data. This is not unexpected as clinical data is seldom well separated, and phenotypes always overlap, making cluster validation difficult. This illuminates the importance, and need for more validation in clustering research. Although, the fourth model was supported by BIC

and SS, it was not an obvious choice as neither metric showed good model fit or separation. Further, the highest relative entropy was at 5 clusters indicating a more certain model, but no other metric indicated 5-cluster model. This model was therefore dismissed as it is often ill advised to solely rely on the peak of the entropy for choosing K.

When assessing the deciding cluster probabilities for the 3-, 4-, and 5-cluster models without LVEF, it becomes apparent that the algorithm is uncertain of which cluster to assign some patients. The cluster probability for the 2-cluster model as shown in the Appendix section A.3.1, is however very left-skewed, indicating a certain clustering. As presented in the result section, the 2-cluster model subdivides into mildly reduced and reduced ejection fraction, while the 3-cluster model splits into two mildly reduced and one reduced group. As there are no HFpEF groups, these models do not entirely support the common classification of HF into HFpEF, HFmrEF, and HFrEF, showing that it is unclear that these groups naturally occur. When clustering with LVEF, there were clinically distinctive preserved and reduced groups while no clusters of HFmrEF were observed, indicating that LVEF plays a crucial role in defining subgroups within heart failure. It also suggests that HFmrEF might not be a distinct phenotype, but rather overlap between HFpEF and HFrEF. Further, when assessing LCA clustered with HF duration, the subsequent clusters were virtually the same as for the main model and HFDIDUR is therefore redundant.

K-prototypes was used to evaluate if the clusters from the main model would be obtained using a different method. This model was useful as it intrinsically handles categorical data, where the Python package automatically tuned the categorical feature weight, preventing the categorical covariates from dominating. The validation metrics indicate that 3 clusters were preferred over 4. The silhouette scores are generally lower in k-prototypes than for LCA, indicating that LCA separates the clusters better, and perhaps that there is overlapping latent distributions that make up the data structure. The clusters in k-prototypes and LCA were similar on the groups with reduced ejection fraction. **Cluster 1**, the youngest mostly male demographic with reduced ejection fraction, clearly corresponds to cluster 4 in the LCA model where it only differs significantly in HBA1C. **Cluster 2** is the second oldest, mostly male group with reduced ejection fraction, whose high symptom burden and long HF duration leads to cachexia. This cluster matches with cluster 2 in the LCA model, where it differs on T2DM, AF, and MI. K-prototypes and LCA, however, differ in their groups with preserved ejection fraction. **Cluster 3** is the oldest group with a preserved ejection fraction, while **Cluster 4** is the other preserved group. These correspond to the LCA clusters 1 and 3 on age, LVEF, and eGFR, but significantly differs other covariates. Although the characteristics of the groups with reduced ejection fraction correspond well between LCA and k-prototypes, they do not necessarily track for the clusters with preserved ejection fraction. Evident from the different cluster sizes in LCA and k-prototypes, restructuring of patients between clusters has likely occurred, resulting in different cluster characteristics. Therefore, k-prototypes does not fully validate the phenotypes observed in LCA for HFpEF.

Several factors should be taken into account simultaneously when interpreting the clustering results. The heterogeneity in study populations and the variations in the inclusion and exclusion criteria of clinical trial data may lead to different latent structures across cohorts. This does not necessarily undermine the validity of findings, but rather reflects meaningful differences in patient characteristics across distinct populations. One should also be mindful of the quality and size of the datasets used for clustering. For example, while registries are often vast, they are often also less regulated and controlled than clinical trial data.

Preprocessing steps, such as if and what type of normalization is performed, influences how much weight features have, depending on their scales. Similarly, based on the method used, feature weighting might be necessary. Therefore, these choices have an impact on clustering, and if one is not careful, it might artificially inflate the importance of some features, leading to erroneous clustering results that do not reflect the underlying data structure. The selection of features is an another important step, and should be done in close collaboration with clinical knowledge to reflect the pathophysiology. In this project, only covariates present across all four trials were included in clustering, leading to the exclusion of otherwise informative features such as KCCQ, COPD, and CAD. Through excluding these or not using additional features, it results in phenotypes that may answer different research questions.

In the feature selection process, T2DM or HBA1C had to be removed due to correlation. HbA1c does not always distinguish diabetic patients well as levels above the diagnostic criteria indicates a high risk of T2DM, but they may not develop it for over 10 years [65]. This, along with BLHBA1C being more right-skewed after log-transformation resulted in retaining T2DM. For SBP and DBP, ablation studies resulted in insignificant differences in cluster quality. While both are important indicators of CV health, SBP was chosen over DBP based on its stronger association with HF events [66]. Hence, retaining T2DM and SBP increases clinical relevance of clustering outcomes.

The choice of clustering algorithm may influence the identification of clinical phenotypes. If, for example, a simple method such as k-means is selected, the features that can be effectively used are limited to numerical covariates. Further, due to the curse of dimensionality, the number of features that one could use is also limited. Many algorithms, such as the ones used in this project, rely on manual selection of the number of clusters. This is often, as in this project, not straightforward due to multiple interpretations of validation metrics. Further, different methods have different strengths and weaknesses where different insights can be obtained from the data. Although data can be preprocessed to allow the application of almost any method, its results would not necessarily reflect real data structures. Moreover, validation metrics containing cohesion and separation are often not applicable to mixed datasets, without using imperfect adaptations, such as Gower's distance. However, because of downstream clinical applications, it is even more important that the validity of clusters is communicated as clearly as possible.

There are a few improvements that could elevate the results of this thesis. First, although this project relied on internal validation metrics, the reproducibility could be indicated over multiple datasets by reproducing the results on a separate validation dataset, or through cross validation. Secondly, removing covariates which did not help clustering could improve model fit and usability. The novel method SHAP (SHapley Additive exPlanations) may support feature selection and enhance clustering results [67].

5
Conclusion

In this project, LCA was used to cluster HF patients based on patient characteristics, typically recorded in HF trials, resulting in the following four clusters:

- Cluster 1: Old AF no MI (HFpEF)
- Cluster 2: Male high NT-ProBNP with MI (HFrEF)
- Cluster 3: Obese diabetic (HFpEF)
- Cluster 4: Young male, with high eGFR (HFrEF)

The HFrEF groups were validated when k-prototypes obtained similar results. These clusters reflect clinical relevance and expected outcomes for those phenotypes. However, these clusters should be validated on external datasets with additional methodologies while maintaining a high quality of cluster validation.

These clusters show poor separation with the validation metrics that were used, either due to limitations in the validation or clustering methods, or reflecting weak patterns in the data. This highlights the importance of further refinement through feature selection.

As there still is no standard for cluster validation methods on mixed clinical data, there is a need for future work in validation methodologies so that the potential for phenotyping and subsequent targeted therapies can be harnessed.

Bibliography

[1] C. D. Kemp and J. V. Conte, "The pathophysiology of heart failure," *Cardiovascular Pathology*, vol. 21, no. 5, pp. 365–371, 2012, ISSN: 1054-8807. DOI: https://doi.org/10.1016/j.carpath.2011.11.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1054880711001529.

[2] G. Savarese, P. M. Becher, L. H. Lund, P. Seferovic, G. M. C. Rosano, and A. J. S. Coats, "Global burden of heart failure: A comprehensive and updated review of epidemiology," en, *Cardiovascular Research*, vol. 118, no. 17, pp. 3272–3287, Jan. 2023, ISSN: 0008-6363, 1755-3245. DOI: 10.1093/cvr/cvac013. [Online]. Available: https://academic.oup.com/cardiovascres/article/118/17/3272/6527627.

[3] T. A. McDonagh, M. Metra, M. Adamo, *et al.*, "2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure," en, *European Heart Journal*, vol. 42, no. 36, pp. 3599–3726, Sep. 2021, ISSN: 0195-668X, 1522-9645. DOI: 10.1093/eurheartj/ehab368. [Online]. Available: https://academic.oup.com/eurheartj/article/42/36/3599/6358045.

[4] R. H. G. Schwinger, "Pathophysiology of heart failure," *Cardiovascular Diagnosis and Therapy*, vol. 11, no. 1, pp. 263–276, Feb. 2021, ISSN: 2223-3652. DOI: 10.21037/cdt-20-302. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7944197/.

[5] T. Ahmad, L. H. Lund, P. Rao, *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," en, *Journal of the American Heart Association*, vol. 7, no. 8, e008081, Apr. 2018, ISSN: 2047-9980. DOI: 10.1161/JAHA.117.008081. [Online]. Available: https://www.ahajournals.org/doi/10.1161/JAHA.117.008081.

[6] C. Meijs, M. L. Handoko, G. Savarese, *et al.*, "Discovering distinct phenotypical clusters in heart failure across the ejection fraction spectrum: A systematic review," en, *Current Heart Failure Reports*, vol. 20, no. 5, pp. 333–349, Oct. 2023, ISSN: 1546-9549. DOI: 10.1007/s11897-023-00615-z. [Online]. Available: https://doi.org/10.1007/s11897-023-00615-z.

[7] A. Uijl, G. Savarese, I. Vaartjes, *et al.*, "Identification of distinct phenotypic clusters in heart failure with preserved ejection fraction," *European Journal of Heart Failure*, vol. 23, no. 6, pp. 973–982, Jun. 2021, ISSN: 1388-9842. DOI: 10.1002/ejhf.2169. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8359985/.

[8] C. C. Aggarwal and C. K. Reddy, *Data clustering: algorithms and applications*. Boca Raton, FL: CRC Press, 2014.

[9] M. S. Khan, M. S. Arshad, S. J. Greene, *et al.*, "Artificial intelligence and heart failure: A state-of-the-art review," en, *European Journal of Heart Failure*, vol. 25, no. 9, pp. 1507–1525, Sep. 2023, ISSN: 1388-9842, 1879-0844. DOI: 10.1002/ejhf.2994. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ejhf.2994.

[10] A. Banerjee, A. Dashtban, S. Chen, *et al.*, "Identifying subtypes of heart failure from three electronic health record sources with machine learning: An external, prognostic, and genetic validation study," en, *The Lancet Digital Health*, vol. 5, no. 6, e370–e379, Jun. 2023, ISSN: 25897500. DOI: 10.1016/S2589-7500(23)00065-1. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2589750023000651.

[11] J. B. Cohen, S. J. Schrauben, L. Zhao, *et al.*, "Clinical phenogroups in heart failure with preserved ejection fraction," en, *JACC: Heart Failure*, vol. 8, no. 3, pp. 172–184, Mar. 2020, ISSN: 22131779. DOI: 10.1016/j.jchf.2019.09.009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2213177919308005.

[12] Å. K. Hedman, C. Hage, A. Sharma, *et al.*, "Identification of novel phenogroups in heart failure with preserved ejection fraction using machine learning," en, *Heart*, vol. 106, pp. 342–349, Mar. 2020, ISSN: 1355-6037, 1468-201X. DOI: 10.1136/heartjnl-2019-315481. [Online]. Available: https://heart.bmj.com/lookup/doi/10.1136/heartjnl-2019-315481.

[13] S. Urban, M. Błaziak, M. Jura, *et al.*, "Novel phenotyping for acute heart failure—unsupervised machine learning-based approach," en, *Biomedicines*, vol. 10, no. 7, p. 1514, Jun. 2022, ISSN: 2227-9059. DOI: 10.3390/biomedicines10071514. [Online]. Available: https://www.mdpi.com/2227-9059/10/7/1514.

[14] F. Schrub, E. Oger, A. Bidaut, *et al.*, "Heart failure with preserved ejection fraction: A clustering approach to a heterogenous syndrome," en, *Archives of Cardiovascular Diseases*, vol. 113, no. 6–7, pp. 381–390, Jun. 2020, ISSN: 18752136. DOI: 10.1016/j.acvd.2020.03.012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1875213620301029.

[15] S. J. Shah, D. H. Katz, S. Selvaraj, *et al.*, "Phenomapping for novel classification of heart failure with preserved ejection fraction," en, *Circulation*, vol. 131, no. 3, pp. 269–279, Jan. 2015, ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIRCULATIONAHA.114.010637. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.114.010637.

[16] L. Monzo, E. Bresso, K. Dickstein, *et al.*, "Machine learning approach to identify phenotypes in patients with ischaemic heart failure with reduced ejection fraction," en, ejhf.3547, Dec. 2024, ISSN: 1388-9842, 1879-0844. DOI: 10.1002/ejhf.3547. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ejhf.3547.

[17] F. Gu, A. J. Meyer, F. Ježek, *et al.*, "Identification of digital twins to guide interpretable ai for diagnosis and prognosis in heart failure," en, *npj Digital Medicine*, vol. 8, no. 1, p. 110, Feb. 2025, ISSN: 2398-6352. DOI: 10.1038/

s41746-025-01501-9. [Online]. Available: https://www.nature.com/articles/s41746-025-01501-9.

[18] C. Martins, B. Neves, A. S. Teixeira, *et al.*, "Identifying subgroups in heart failure patients with multimorbidity by clustering and network analysis," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 95, Apr. 2024, ISSN: 1472-6947. DOI: 10.1186/s12911-024-02497-0. [Online]. Available: https://doi.org/10.1186/s12911-024-02497-0.

[19] D. P. Kao, J. D. Lewsey, I. S. Anand, *et al.*, "Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response," en, *European Journal of Heart Failure*, vol. 17, no. 9, pp. 925–935, Sep. 2015, ISSN: 1388-9842, 1879-0844. DOI: 10.1002/ejhf.327. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ejhf.327.

[20] A. Jasinska-Piadlo, R. Bond, P. Biglarbeigi, *et al.*, "Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset," en, *International Journal of Data Science and Analytics*, vol. 15, no. 1, pp. 49–66, Jan. 2023, ISSN: 2364-415X, 2364-4168. DOI: 10.1007/s41060-022-00346-9. [Online]. Available: https://link.springer.com/10.1007/s41060-022-00346-9.

[21] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques* (Morgan Kaufmann series in data management systems), eng, 3rd ed. Amsterdam Boston: Elsevier/Morgan Kaufmann, 2012, ISBN: 9780123814791.

[22] *Regulations for the use of AI tools in thesis work*, en, Dec. 2023. [Online]. Available: https://www.chalmers.se/en/education/your-studies/masters-and-bachelors-thesis/regulations-for-the-use-of-ai-tools/ (visited on 05/18/2025).

[23] OpenAI, *Chatgpt (may 2025 version)*, https://chat.openai.com, Generative AI language model used for language assistance and idea refinement, 2025.

[24] D. Saeed, T. Reza, M. W. Shahzad, *et al.*, "Navigating the crossroads: Understanding the link between chronic kidney disease and cardiovascular health," *Cureus*, vol. 15, no. 12, e51362, ISSN: 2168-8184. DOI: 10.7759/cureus.51362. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10825078/.

[25] J. M. Ter Maaten, K. Damman, M. C. Verhaar, *et al.*, "Connecting heart failure with preserved ejection fraction and renal dysfunction: The role of endothelial dysfunction and inflammation," en, *European Journal of Heart Failure*, vol. 18, no. 6, pp. 588–598, Jun. 2016, ISSN: 1388-9842, 1879-0844. DOI: 10.1002/ejhf.497. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ejhf.497.

[26] P. P. Toth, *English: HFpEF and HFrEF primarily differ primarily in terms of ejection fraction (EF) among other aspects*. Nov. 2020. [Online]. Available: https://commons.wikimedia.org/wiki/File:Key_Differences_Between_HFpEF_and_HFrEF.png.

[27] N. Marx, M. Federici, K. Schütt, *et al.*, "2023 esc guidelines for the management of cardiovascular disease in patients with diabetes," en, *European Heart Journal*, vol. 44, no. 39, pp. 4043–4140, Oct. 2023. DOI: 10.1093/eurheartj/

ehad192. [Online]. Available: https://academic.oup.com/eurheartj/article/44/39/4043/7238227.

[28] R. H. Ritchie and E. D. Abel, "Basic mechanisms of diabetic heart disease," en, *Circulation Research*, vol. 126, no. 11, pp. 1501–1525, May 2020, ISSN: 0009-7330, 1524-4571. DOI: 10.1161/CIRCRESAHA.120.315913. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.120.315913.

[29] B. A. Borlaug, M. D. Jensen, D. W. Kitzman, C. S. P. Lam, M. Obokata, and O. J. Rider, "Obesity and heart failure with preserved ejection fraction: New insights and pathophysiological targets," *Cardiovascular Research*, vol. 118, no. 18, pp. 3434–3450, Jul. 2022, ISSN: 0008-6363. DOI: 10.1093/cvr/cvac120. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10202444/.

[30] en, Oct. 2018. [Online]. Available: https://www.nhs.uk/conditions/coronary-heart-disease/.

[31] M. Gheorghiade, G. Sopko, L. De Luca, *et al.*, "Navigating the crossroads of coronary artery disease and heart failure," en, *Circulation*, vol. 114, no. 11, pp. 1202–1213, Sep. 2006, ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIRCULATIONAHA.106.623199. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.106.623199.

[32] D. V. M. Verhaert, H.-P. Brunner-La Rocca, D. J. van Veldhuisen, and K. Vernooy, "The bidirectional interaction between atrial fibrillation and heart failure: Consequences for the management of both diseases," *Europace*, vol. 23, no. Suppl 2, pp. ii40–ii45, Apr. 2021, ISSN: 1099-5129. DOI: 10.1093/europace/euaa368. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8035705/.

[33] R. Gopinathannair, L. Y. Chen, M. K. Chung, *et al.*, "Managing atrial fibrillation in patients with heart failure and reduced ejection fraction: A scientific statement from the american heart association," en, *Circulation: Arrhythmia and Electrophysiology*, vol. 14, no. 7, Jul. 2021, ISSN: 1941-3149, 1941-3084. DOI: 10.1161/HAE.0000000000000078. [Online]. Available: https://www.ahajournals.org/doi/10.1161/HAE.0000000000000078.

[34] V. Kittipibul and C. S. P. Lam, "Heart failure with preserved ejection fraction and atrial fibrillation: Epidemiology, pathophysiology, and diagnosis interplay," en, *Heart Failure Reviews*, Jan. 2025, ISSN: 1573-7322. DOI: 10.1007/s10741-025-10488-0. [Online]. Available: https://doi.org/10.1007/s10741-025-10488-0.

[35] J. A. Regan, D. W. Kitzman, E. S. Leifer, *et al.*, "Impact of age on comorbidities and outcomes in heart failure with reduced ejection fraction," *JACC. Heart failure*, vol. 7, no. 12, pp. 1056–1065, Dec. 2019, ISSN: 2213-1779. DOI: 10.1016/j.jchf.2019.09.004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6936107/.

[36] X. Huo, B. Pu, W. Wang, *et al.*, "New york heart association class and kansas city cardiomyopathy questionnaire in acute heart failure," *JAMA Network Open*, vol. 6, no. 10, e2339458, Oct. 2023, ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2023.39458. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10599126/.

[37] M. Yang, T. Kondo, C. Adamson, *et al.*, "Impact of comorbidities on health status measured using the kansas city cardiomyopathy questionnaire in patients with heart failure with reduced and preserved ejection fraction," eng, *European Journal of Heart Failure*, vol. 25, no. 9, pp. 1606–1618, Sep. 2023, ISSN: 1879-0844. DOI: 10.1002/ejhf.2962.

[38] K. Kirchner, J. Zec, and B. Delibašić, "Facilitating data preprocessing by a generic framework: A proposal for clustering," en, *Artificial Intelligence Review*, vol. 45, no. 3, pp. 271–297, Mar. 2016, ISSN: 1573-7462. DOI: 10.1007/s10462-015-9446-6. [Online]. Available: https://doi.org/10.1007/s10462-015-9446-6 (visited on 05/07/2025).

[39] C. K. Enders, *Applied missing data analysis*. Guilford Publications, 2022.

[40] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?" en, *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, Mar. 2011, ISSN: 1049-8931, 1557-0657. DOI: 10.1002/mpr.329. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/mpr.329 (visited on 05/14/2025).

[41] P. Sinha, C. S. Calfee, and K. L. Delucchi, "Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls," en, *Critical Care Medicine*, vol. 49, no. 1, e63–e79, Jan. 2021, ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000004710. [Online]. Available: https://journals.lww.com/10.1097/CCM.0000000000004710 (visited on 02/03/2025).

[42] M. Riani, A. C. Atkinson, and A. Corbellini, "Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression," en, *Statistical Methods & Applications*, vol. 32, no. 1, pp. 75–102, Mar. 2023, ISSN: 1613-981X. DOI: 10.1007/s10260-022-00640-7. [Online]. Available: https://doi.org/10.1007/s10260-022-00640-7 (visited on 05/13/2025).

[43] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*, Springer, 2001, pp. 420–434.

[44] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 257–258, ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_192. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_192.

[45] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," en, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, Apr. 2016, ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2015.0202. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 (visited on 05/12/2025).

[46] B. Rafieian, P. Hermosilla, and P.-P. Vázquez, "Improving Dimensionality Reduction Projections for Data Visualization," en, *Applied Sciences*, vol. 13, no. 17, p. 9967, Sep. 2023, ISSN: 2076-3417. DOI: 10.3390/app13179967. [Online]. Available: https://www.mdpi.com/2076-3417/13/17/9967 (visited on 05/18/2025).

[47] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 1200–1205. DOI: 10.1109/MIPRO.2015.7160458.

[48] D.-T. Dinh, V.-N. Huynh, and S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Information Sciences*, vol. 571, pp. 418–442, Sep. 2021, ISSN: 0020-0255. DOI: 10.1016/j.ins.2021.04.076. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025521004114.

[49] Z. Jia and L. Song, "Weighted k-prototypes clustering algorithm based on the hybrid dissimilarity coefficient," en, *Mathematical Problems in Engineering*, vol. 2020, pp. 1–13, Jul. 2020, ISSN: 1024-123X, 1563-5147. DOI: 10.1155/2020/5143797. [Online]. Available: https://www.hindawi.com/journals/mpe/2020/5143797/.

[50] P. Liu, H. Yuan, Y. Ning, B. Chakraborty, N. Liu, and M. A. Peres, "A modified and weighted gower distance-based clustering analysis for mixed type data: A simulation and empirical analyses," *BMC Medical Research Methodology*, vol. 24, no. 1, p. 305, Dec. 2024, ISSN: 1471-2288. DOI: 10.1186/s12874-024-02427-8. [Online]. Available: https://doi.org/10.1186/s12874-024-02427-8.

[51] A. Pyae, Y.-C. Low, and H. N. Chua, "A combined distance metric approach with weight adjustment for improving mixed data clustering quality," in *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, Aug. 2024, pp. 183–188. DOI: 10.1109/IICAIET62352.2024.10730392. [Online]. Available: https://ieeexplore.ieee.org/document/10730392.

[52] G. McLachlan and D. Peel, *Finite Mixture Models* (Wiley Series in Probability and Statistics), en, 1st ed. Wiley, Sep. 2000, ISBN: 9780471006268 9780471721185. DOI: 10.1002/0471721182. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/0471721182 (visited on 05/13/2025).

[53] *English: A replacement for Gaussian-mixture-example.png*, Jun. 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:Gaussian-mixture-example.svg (visited on 05/18/2025).

[54] M. Goller, E. Kyndt, S. Paloniemi, and C. Damsa, Eds., *Methods for researching professional learning and development: challenges, applications and empirical illustrations* (Professional and practice-based learning 33), eng. Cham: Springer International Publishing AG, 2022, ISBN: 9783031085178 9783031085185.

[55] S. Morin, R. Legault, F. Laliberté, *et al.*, "Stepmix: A python package for pseudo-likelihood estimation of generalized mixture models with external variables," *arXiv preprint arXiv:2304.03853*, 2023.

[56] Ö. Akay and G. Yüksel, "Clustering the mixed panel dataset using gower's distance and k-prototypes algorithms," en, *Communications in Statistics - Simulation and Computation*, vol. 47, no. 10, pp. 3031–3041, Nov. 2018, ISSN: 0361-0918, 1532-4141. DOI: 10.1080/03610918.2017.1367806. [Online]. Available:

https://www.tandfonline.com/doi/full/10.1080/03610918.2017.1367806.

[57] N. J. de Vos, *Kmodes categorical clustering library*, https://github.com/nicodv/kmodes, 2015–2024.

[58] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," en, *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998, ISSN: 1573-756X. DOI: 10.1023/A:1009769707641. [Online]. Available: https://doi.org/10.1023/A:1009769707641.

[59] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved k-prototypes clustering algorithm for mixed numeric and categorical data," en, *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013, ISSN: 09252312. DOI: 10.1016/j.neucom.2013.04.011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231213004773 (visited on 05/18/2025).

[60] C. X. Gao, D. Dwyer, Y. Zhu, *et al.*, "An overview of clustering methods with guidelines for application in mental health research," *Psychiatry Research*, vol. 327, p. 115 265, Sep. 2023, ISSN: 0165-1781. DOI: 10.1016/j.psychres.2023.115265. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165178123002159.

[61] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic)," *Psychological Methods*, vol. 17, no. 2, pp. 228–243, Jun. 2012, ISSN: 1082-989X. DOI: 10.1037/a0027127. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3366160/.

[62] A. Hoayek and D. Rullière, "Assessing clustering methods using shannon's entropy," *Information Sciences*, vol. 689, p. 121 510, Jan. 2025, ISSN: 0020-0255. DOI: 10.1016/j.ins.2024.121510. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025524014245.

[63] K. L. Nylund, T. Asparouhov, and B. O. Muthén, "Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study," en, *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 14, no. 4, pp. 535–569, Oct. 2007, ISSN: 1070-5511, 1532-8007. DOI: 10.1080/10705510701575396. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10705510701575396.

[64] T. Ullmann, C. Hennig, and A.-L. Boulesteix, "Validation of cluster analysis results on validation data: A systematic framework," en, *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 3, e1444, May 2022, ISSN: 1942-4787, 1942-4795. DOI: 10.1002/widm.1444. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1444.

[65] A. D. Pradhan, N. Rifai, J. E. Buring, and P. M. Ridker, "Hemoglobin A1c Predicts Diabetes but Not Cardiovascular Disease in Nondiabetic Women," en, *The American Journal of Medicine*, vol. 120, no. 8, pp. 720–727, Aug. 2007, ISSN: 00029343. DOI: 10.1016/j.amjmed.2007.03.022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0002934307004524 (visited on 05/15/2025).

[66] A. W. Haider, M. G. Larson, S. S. Franklin, and D. Levy, "Systolic Blood Pressure, Diastolic Blood Pressure, and Pulse Pressure as Predictors of Risk for Congestive Heart Failure in the Framingham Heart Study," en, *Annals of Internal Medicine*, vol. 138, no. 1, pp. 10–16, Jan. 2003, ISSN: 0003-4819, 1539-3704. DOI: 10.7326/0003-4819-138-1-200301070-00006. [Online]. Available: https://www.acpjournals.org/doi/10.7326/0003-4819-138-1-200301070-00006 (visited on 05/18/2025).

[67] W. E. Marcílio-Jr and D. M. Eler, "Explaining dimensionality reduction results using shapley values," *Expert Systems with Applications*, vol. 178, p. 115 020, Sep. 2021, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2021.115020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421004619.

A

Appendix 1

The Appendix presents the results for 3-, 4- and 5-cluster models which were not presented in the main report. These results include the covariate distributions per cluster in the form of histograms and heatmaps along with survival curves for models trained on the main LCA model, LCA with HF duration, LCA without LVEF, and k-Prototypes. Additionally, the probability distributions per cluster is presented for the 2-cluster LCA models.

A.1 Latent Class Analysis

Relevant results for the LCA models are presented that were not included in the main report.

A.1.1 Initial Visualizations

When the models were clustered, the clusters were first visualized using PCA for the most important models from 2- to 5-clusters, shown in Figure A.1.

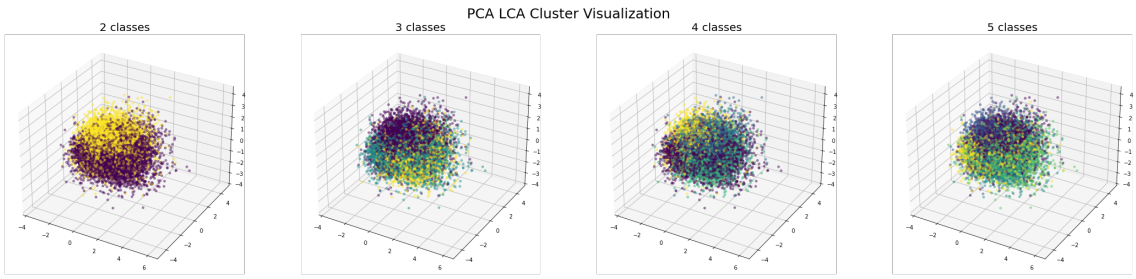

Figure A.1: Scatter plot of LCA 2- to 5-cluster models using PCA.

A.1.2 2-Cluster Model

In Figure A.2, the deciding probabilities for each patient are presented in a histogram per cluster to evaluate the certainty of the clustering.

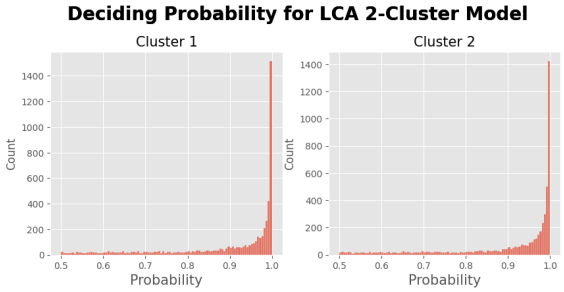

Figure A.2: LCA 2-cluster model deciding probabilities.

A.1.3 3-Cluster Model

Figure A.3 shows the histogram distributions per covariate across all datasets for the 3-cluster LCA model, post-clustering.

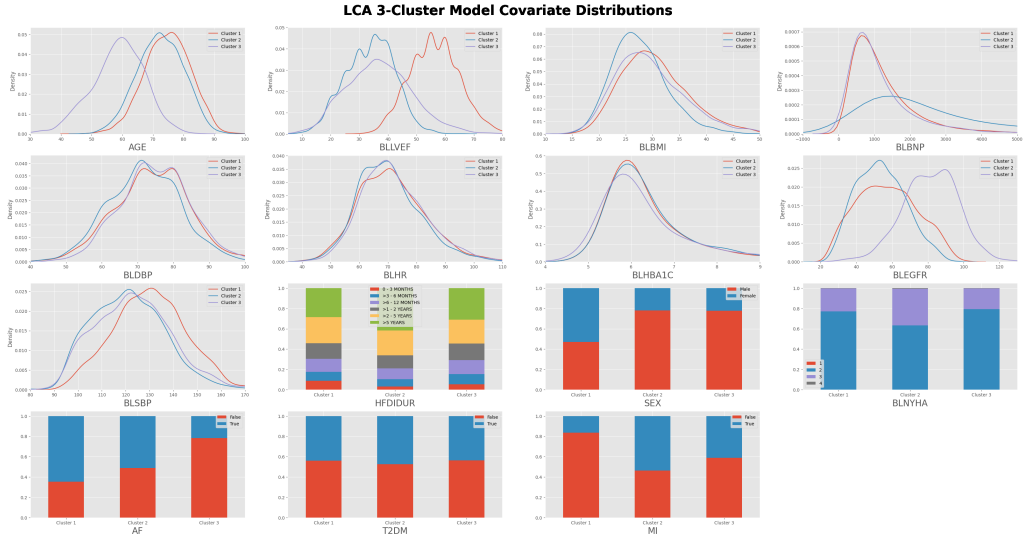

Figure A.3: Covariate histograms for per dataset for the LCA 3 cluster model

In Figure A.4, the dichotomized covariate heatmap is presented for the LCA 3-cluster model.

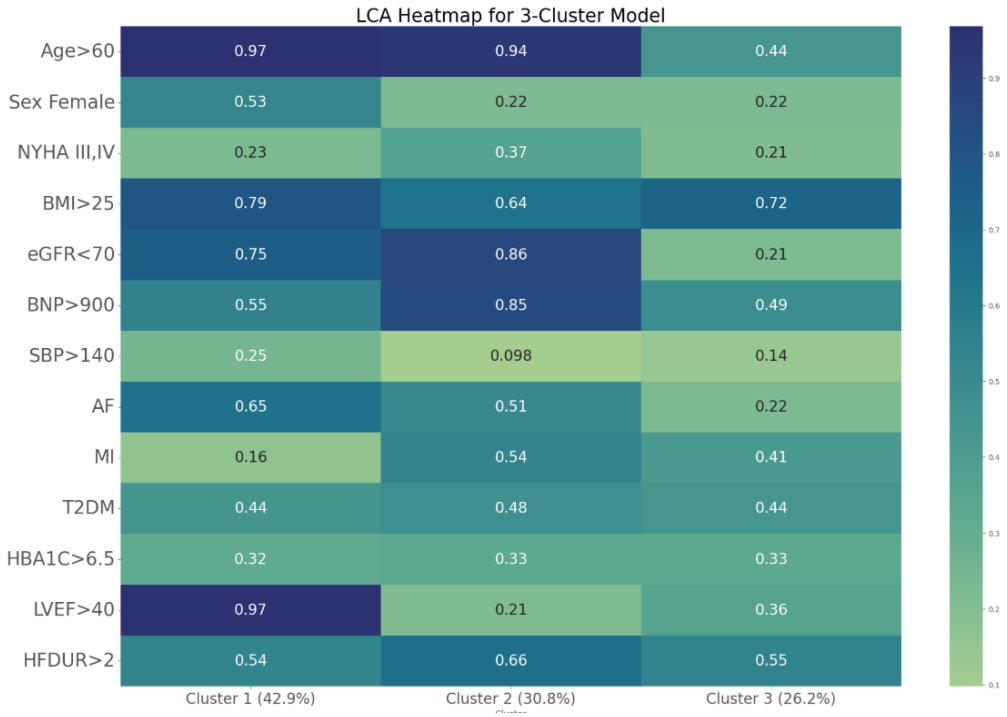

Figure A.4: Heatmap for the LCA 3 cluster model

A.1.4 4-Cluster Model

Figure A.5, shows the dichotomized covariate heatmap for the LCA 4-cluster model.

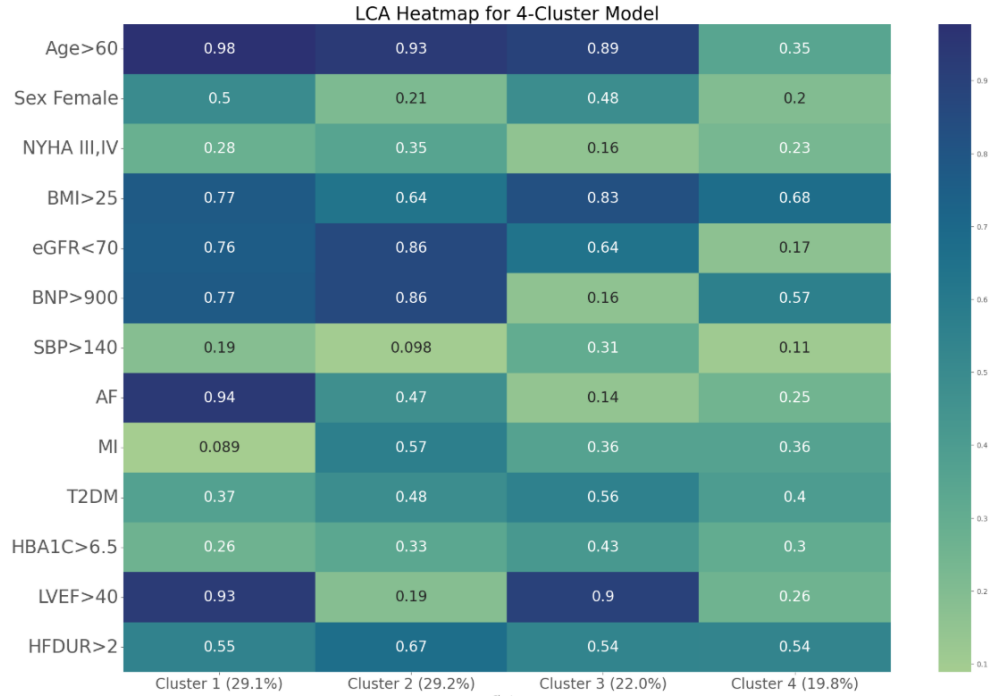

Figure A.5: Heatmap for the LCA 4-cluster model

A.1.5 5-Cluster Model

Figure A.6 presents the covariate distributions per cluster for the 5-cluster LCA model.

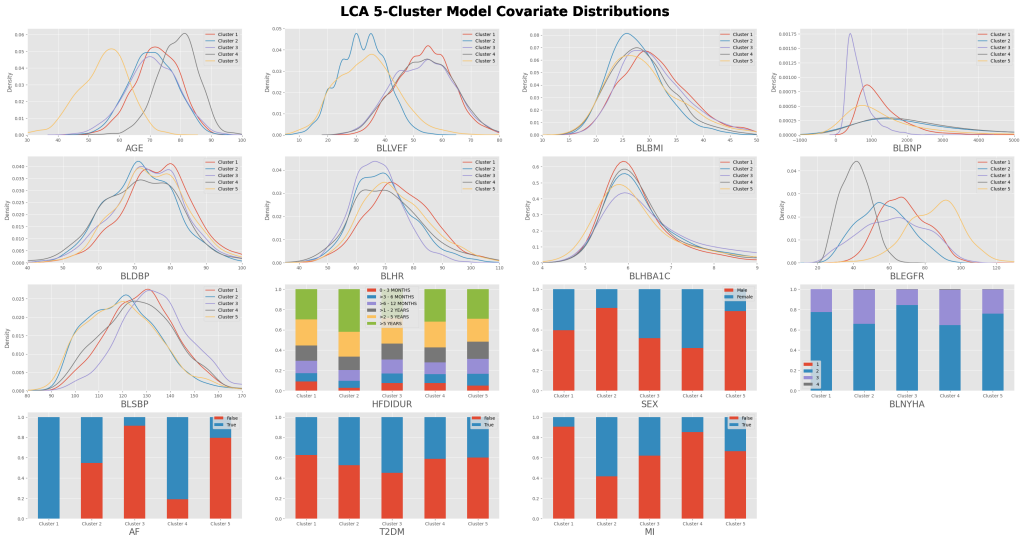

Figure A.6: Covariate histograms per cluster for the LCA 5 cluster model

In Figure A.7, the resulting heatmap for the 5-cluster LCA model is presented.

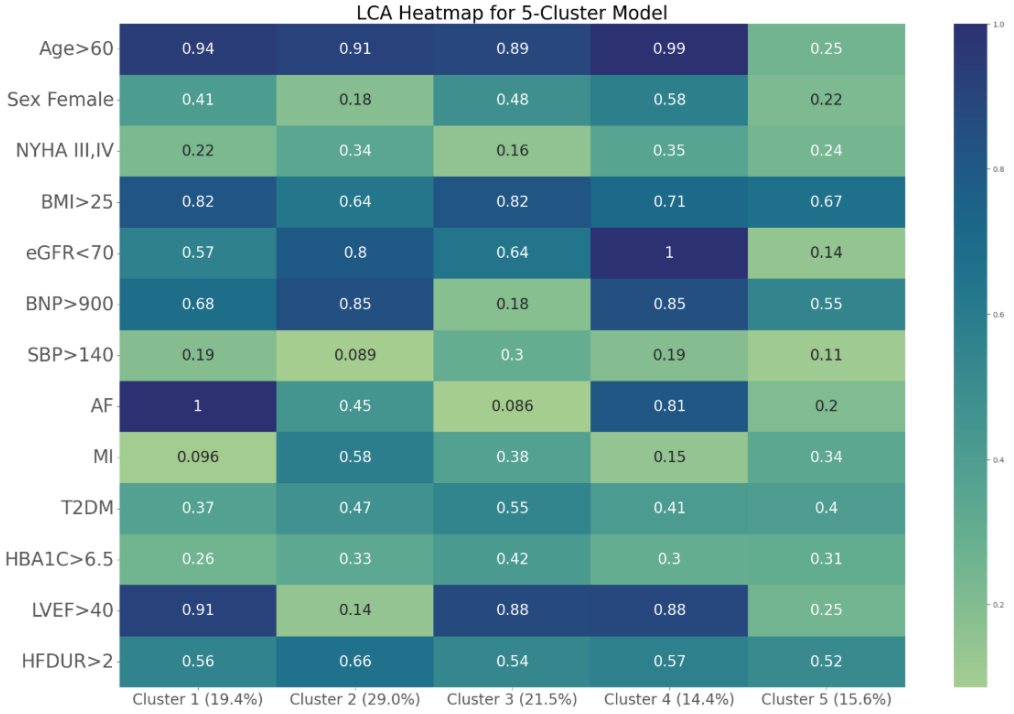

Figure A.7: Heatmap for the LCA 5 cluster model

A.2 LCA with HF Duration

The results from training LCA when heart failure duration (HFDIDUR) was added to the clustering covariates are presented in this section.
Figure A.8 shows the BIC scores, relative entropies, and silhouette scores for the 2- to 10-cluster models.

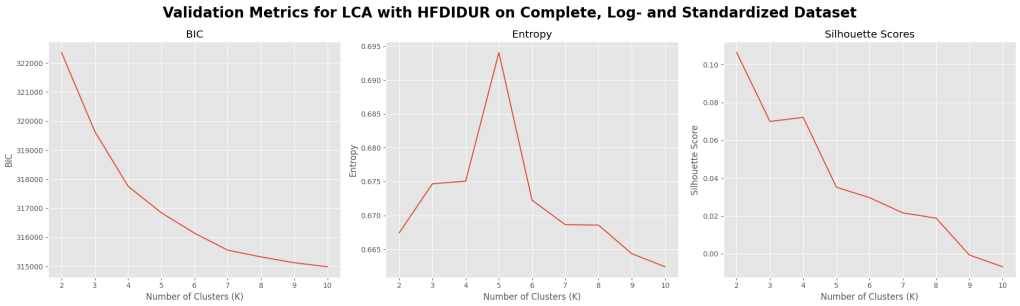

Figure A.8: Validation Metrics for LCA with HF Duration

Figure A.9 shows the deciding probabilities for the LCA 2-, 3-, 4-, and 5-cluster models.

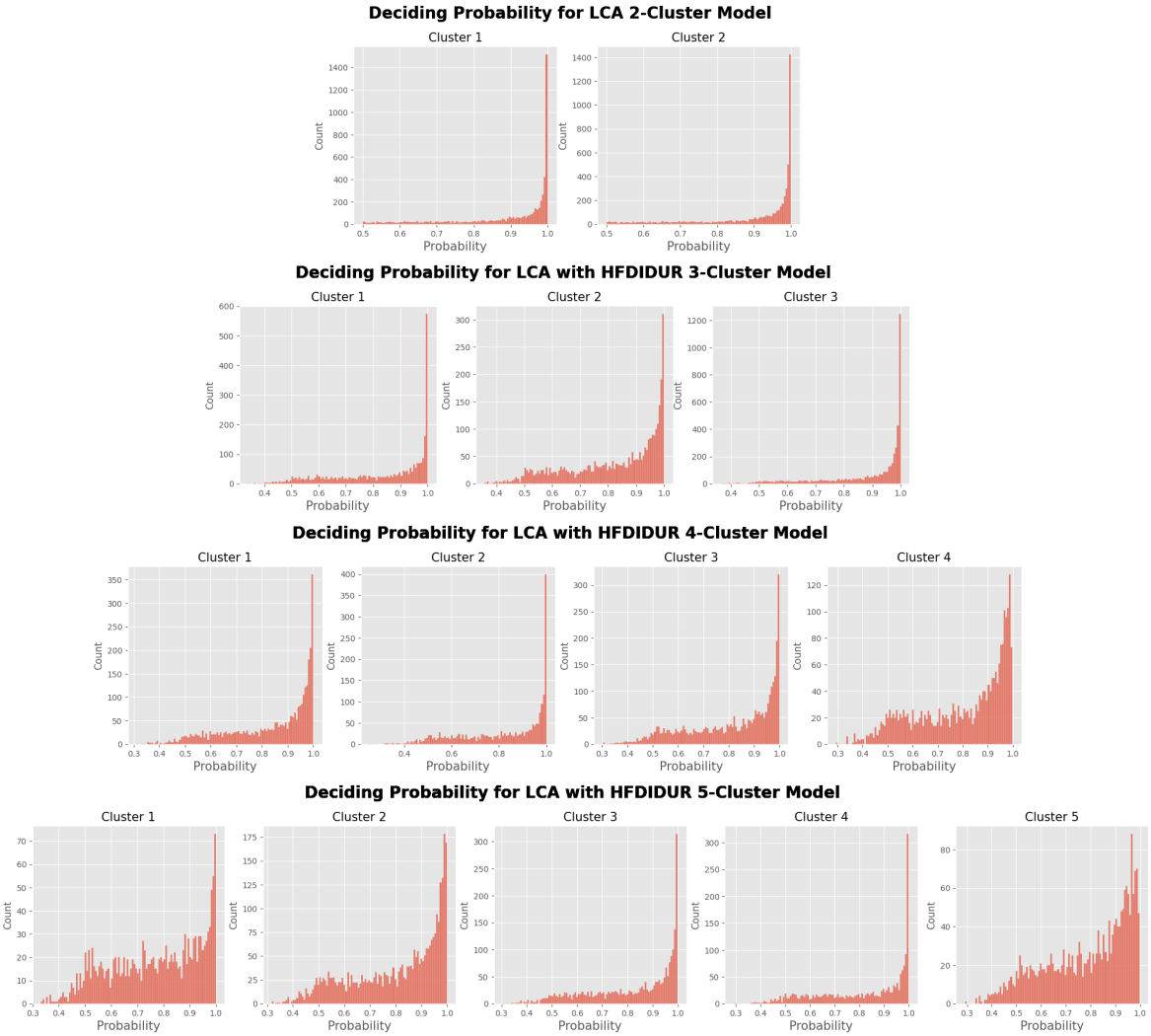

Figure A.9: The deciding probabilities for the LCA model with HF duration for 3, 4, and 5 clusters.

A.2.1 3-Cluster Model

Figure A.10 shows the histogram distributions per covariate across all datasets for the 3-cluster LCA model clustered with heart failure duration.

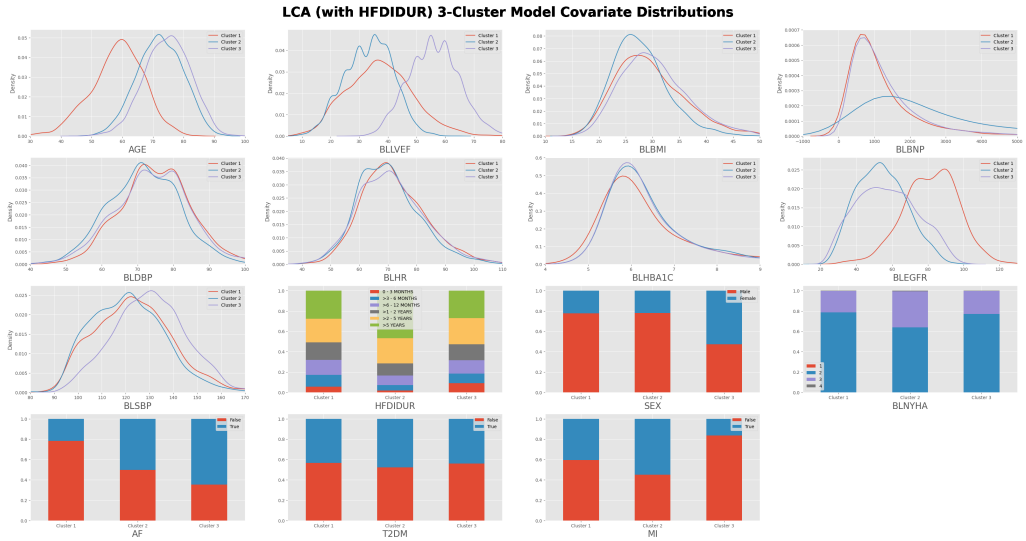

Figure A.10: Covariate histograms for per dataset for the LCA 3 cluster model with HF Duration

In Figure A.11, the dichotomized covariate heatmap is presented for the LCA 3-cluster model with heart failure duration.

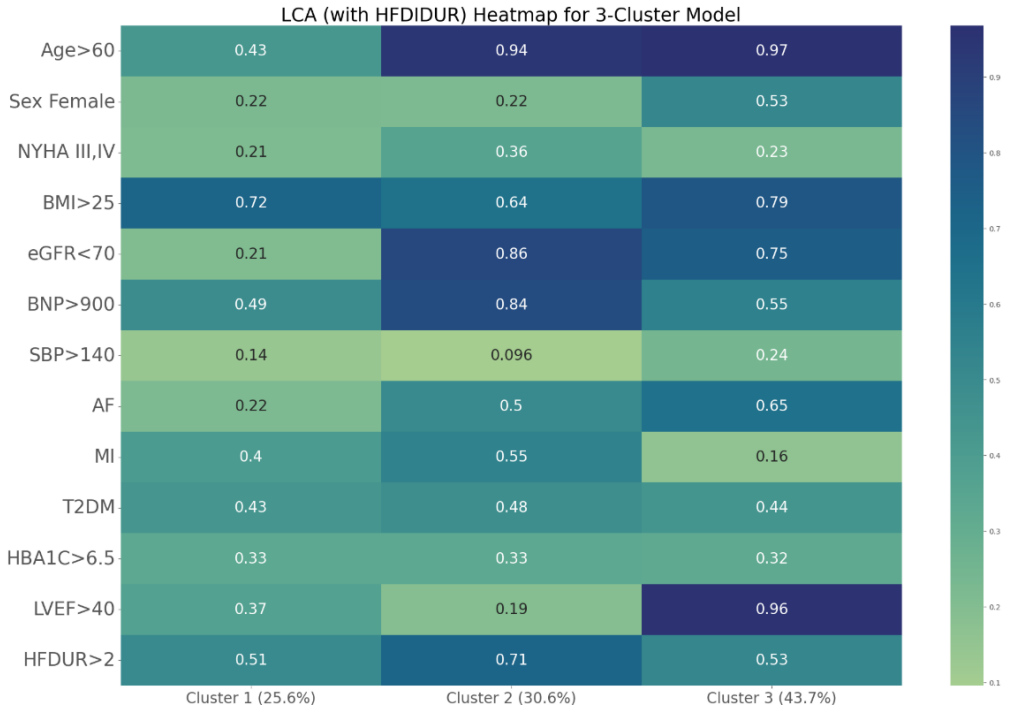

Figure A.11: Heatmap for the LCA 3 cluster model with HF Duration

A.2.2 4-Cluster Model

Figure A.12 shows the covariate distribution per cluster for the LCA 4-cluster model with HF Duration.

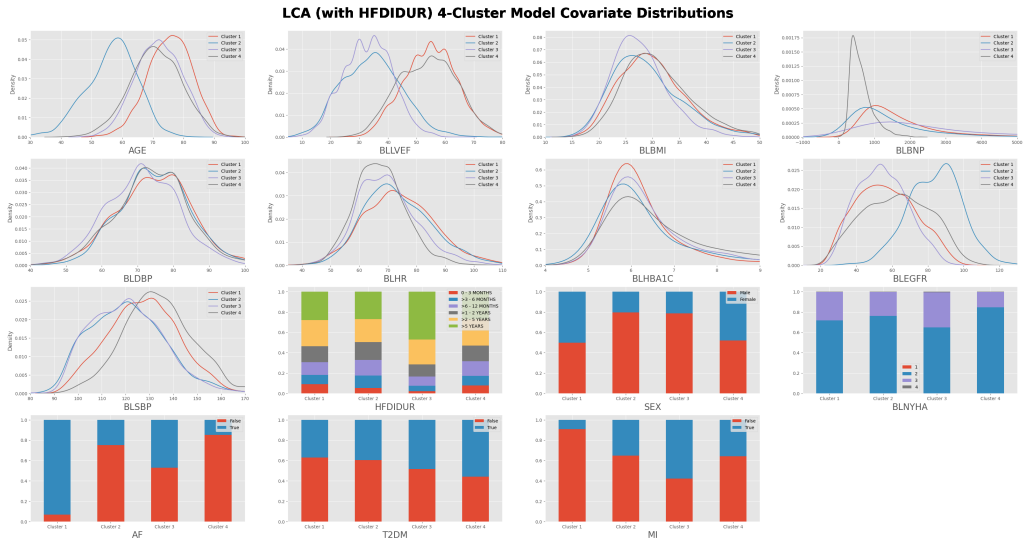

Figure A.12: Covariate histograms per cluster for the LCA 4 cluster model with HF Duration

In Figure A.13, the heatmap for the LCA 4-cluster model with HF Duration is presented.

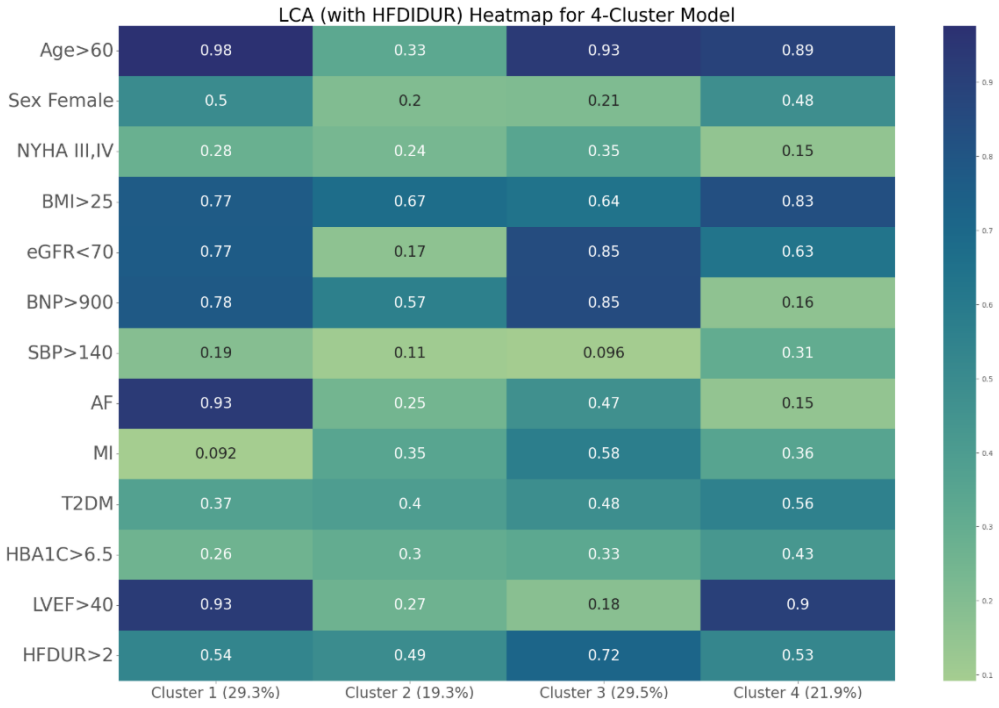

Figure A.13: Heatmap for the LCA 5 cluster model with HF Duration

A.2.3 5-Cluster Model

Figure A.14 shows the covariate distribution per cluster for the LCA 5-cluster model with HF Duration.

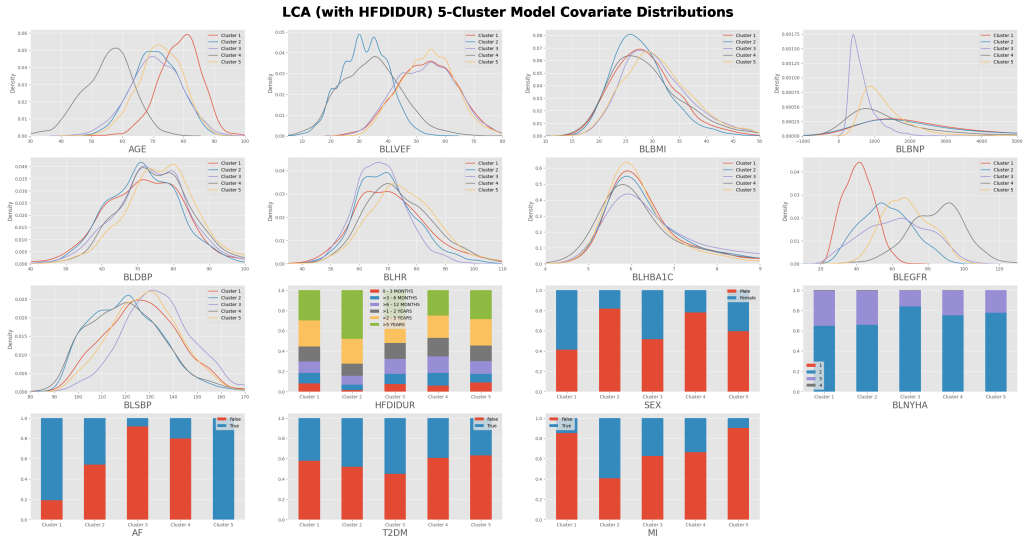

Figure A.14: Covariate histograms per cluster for the LCA 5 cluster model with HF Duration

In Figure A.15, the heatmap for the LCA 5-cluster model with HF Duration is presented.

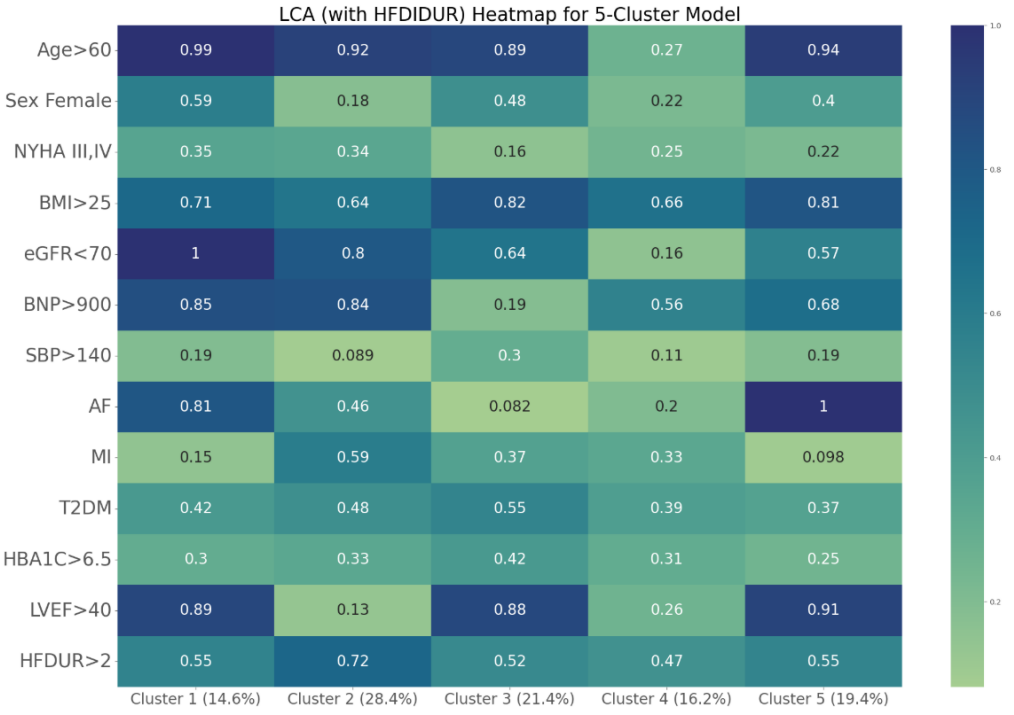

Figure A.15: Heatmap for the LCA 5 cluster model with HF Duration

A.3 LCA without LVEF

This section presents all relevant results for LCA clustered without LVEF that were not presented in the main report.

A.3.1 2-Cluster Model

Figure A.16 show the desciding probabilities for each cluster in LCA 2-cluster model without LVEF.

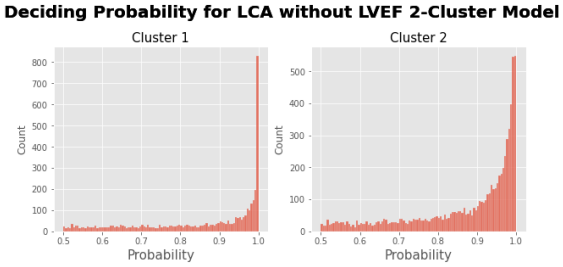

Figure A.16: Deciding probabilities for LCA 2-cluster model with LVEF.

A.3.2 3-Cluster Model

Figure A.17 shows the covariate distribution per cluster for the LCA 3-cluster model without LVEF.

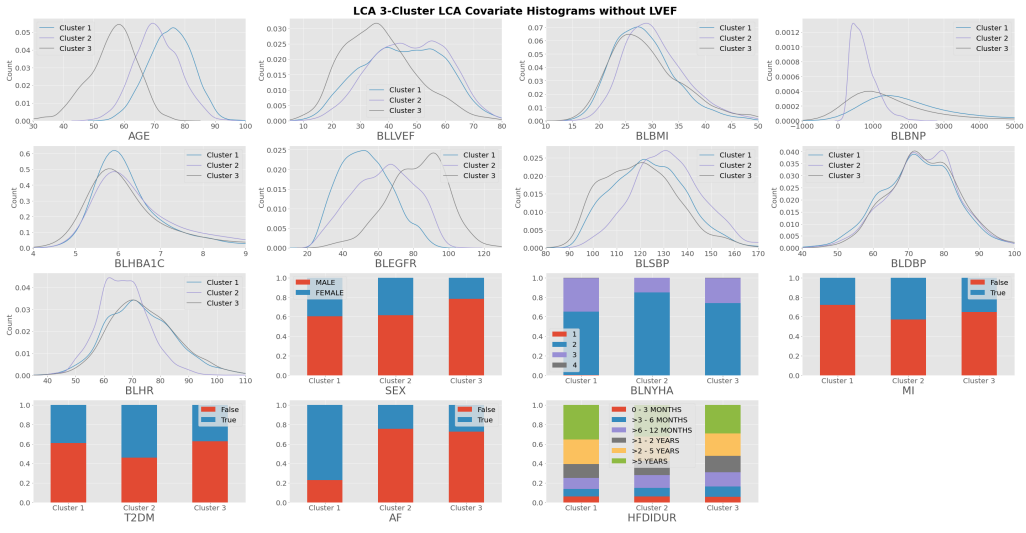

Figure A.17: Covariate histograms per cluster for the LCA 3 cluster model without LVEF

Figure A.18 shows the heatmap for the covariates for the 3-cluster LCA model without LVEF.

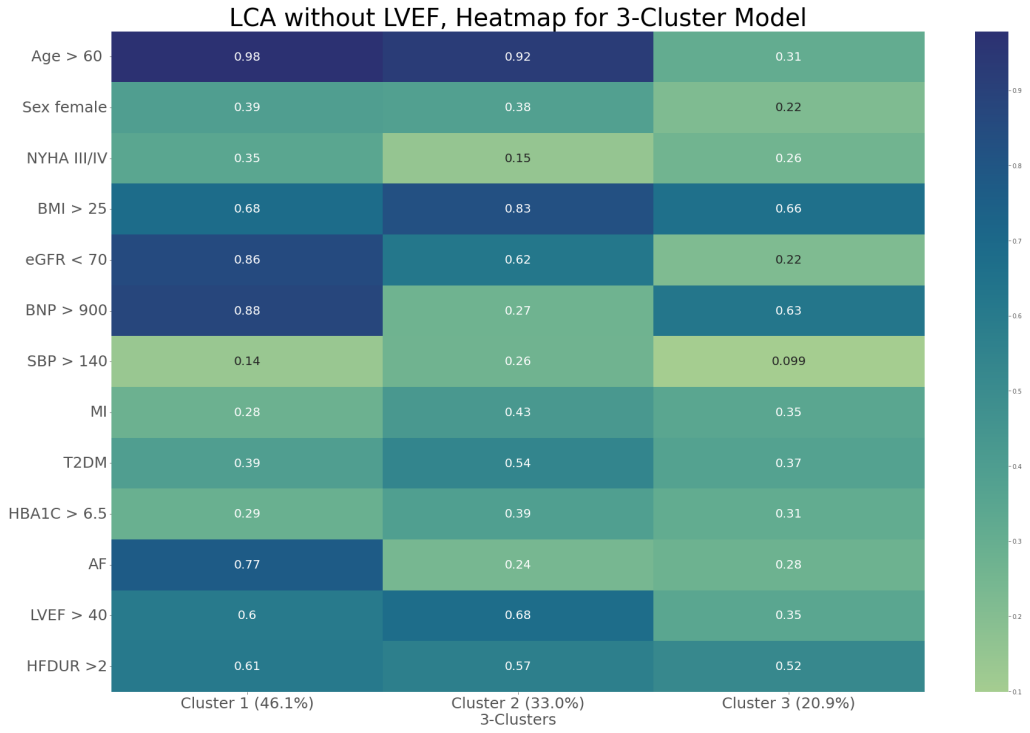

Figure A.18: Heatmap for the LCA 3 cluster model without LVEF

A.3.3 4-Cluster Model

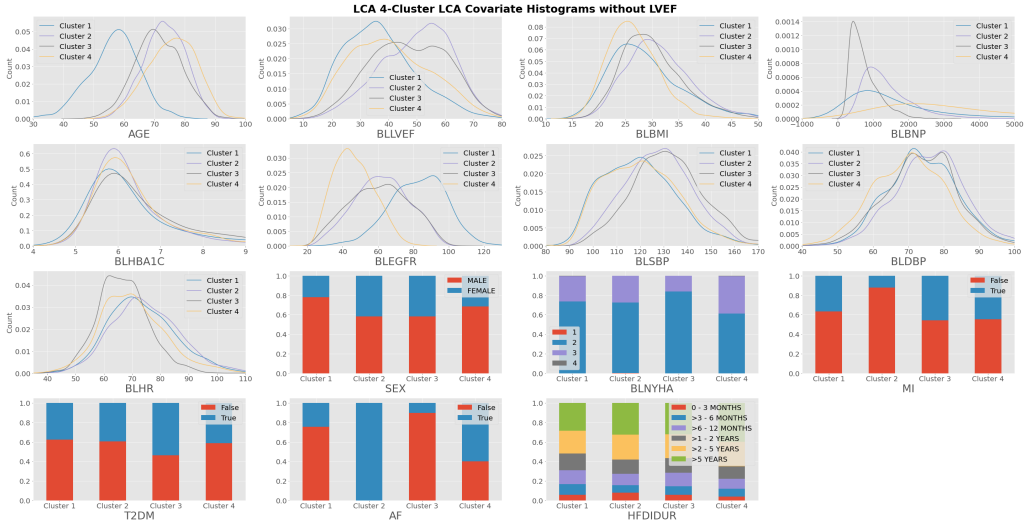

Figure A.19: Covariate histograms per cluster for the LCA 4 cluster model without LVEF

In Figure A.20, the heatmap for the covariates for the 4-cluster LCA model without LVEF is displayed.

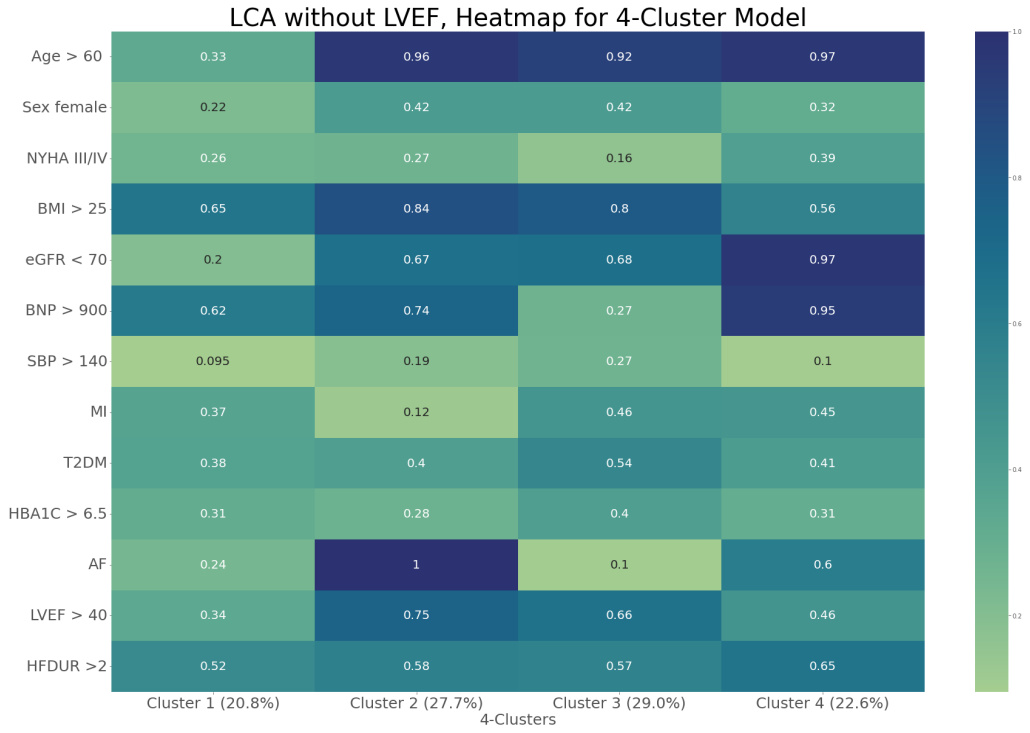

Figure A.20: Heatmap for the LCA 4 cluster model without LVEF

A.3.4 5-Cluster Model

Figure A.21 shows the covariate distribution per cluster for the LCA 5-cluster model without LVEF.

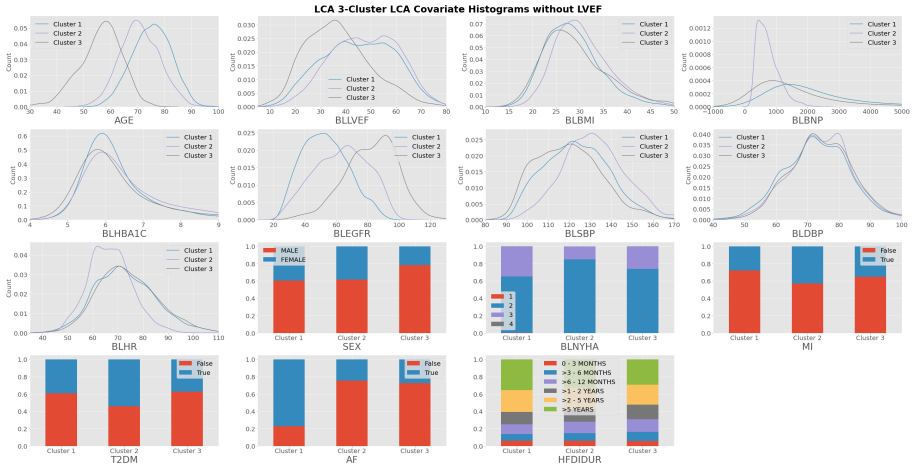

Figure A.21: Covariate histograms per cluster for the LCA 3 cluster model without LVEF

Figure A.22 shows the heatmap for the covariates for the 5-cluster LCA model without LVEF.

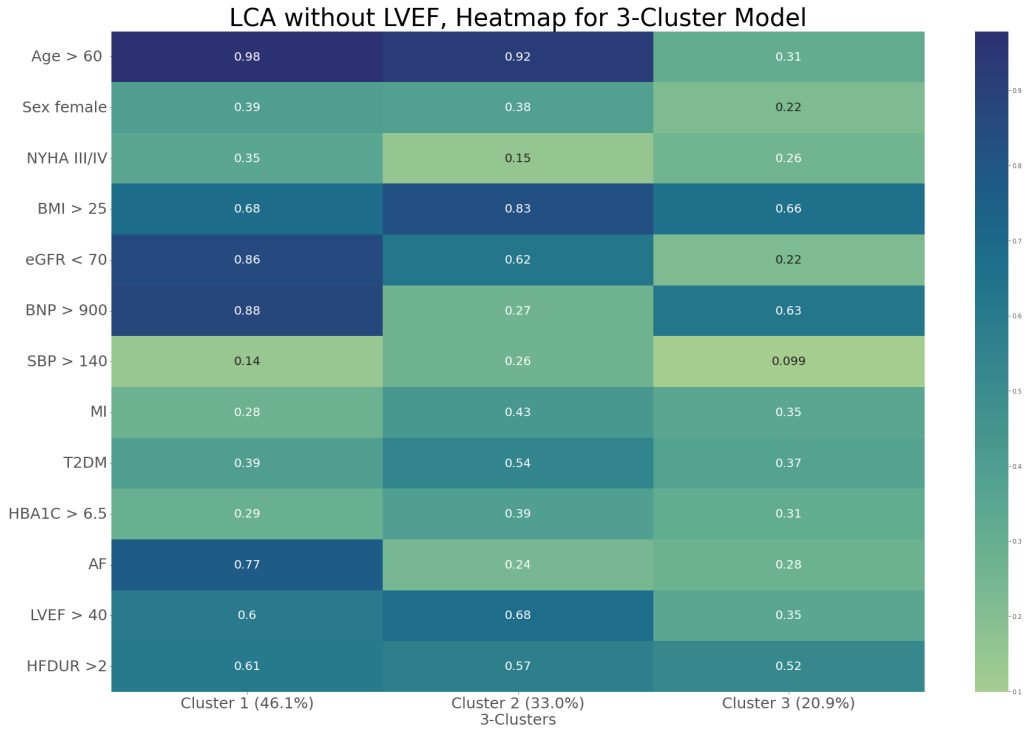

Figure A.22: Heatmap for the LCA 3 cluster model without LVEF

A.4 K-Prototypes

This section presents the result for the 3-, 4- and 5-cluster k-Prototype model. First, Figure A.23 shows the dichotomized covariate heatmap for the 3-cluster model.

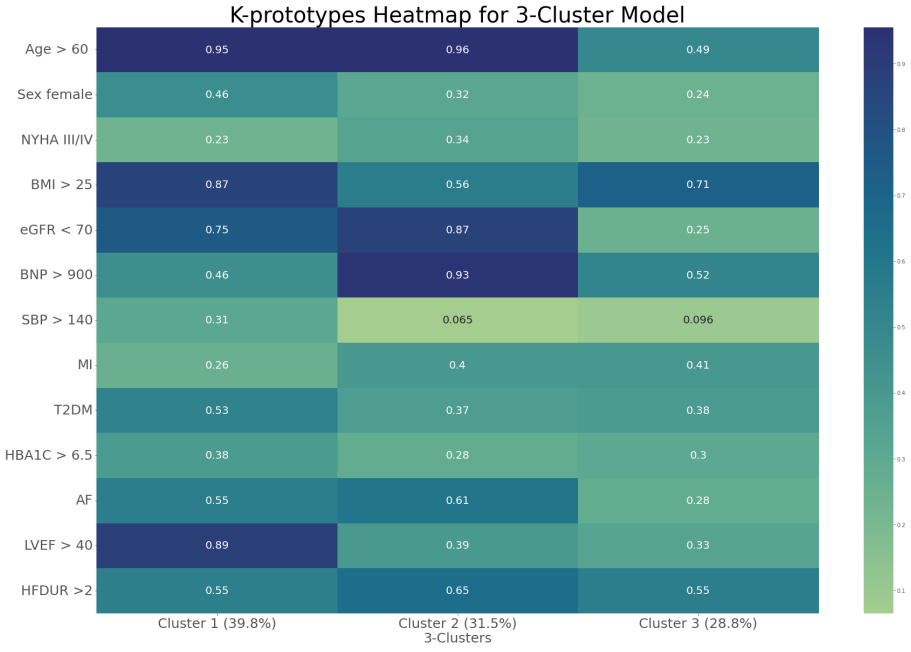

Figure A.23: Heatmap for the 3-cluster k-Prototype model

Secondly, Figure A.24 shows the dichotomized heatmap for the 4-cluster k-Prototypes model.

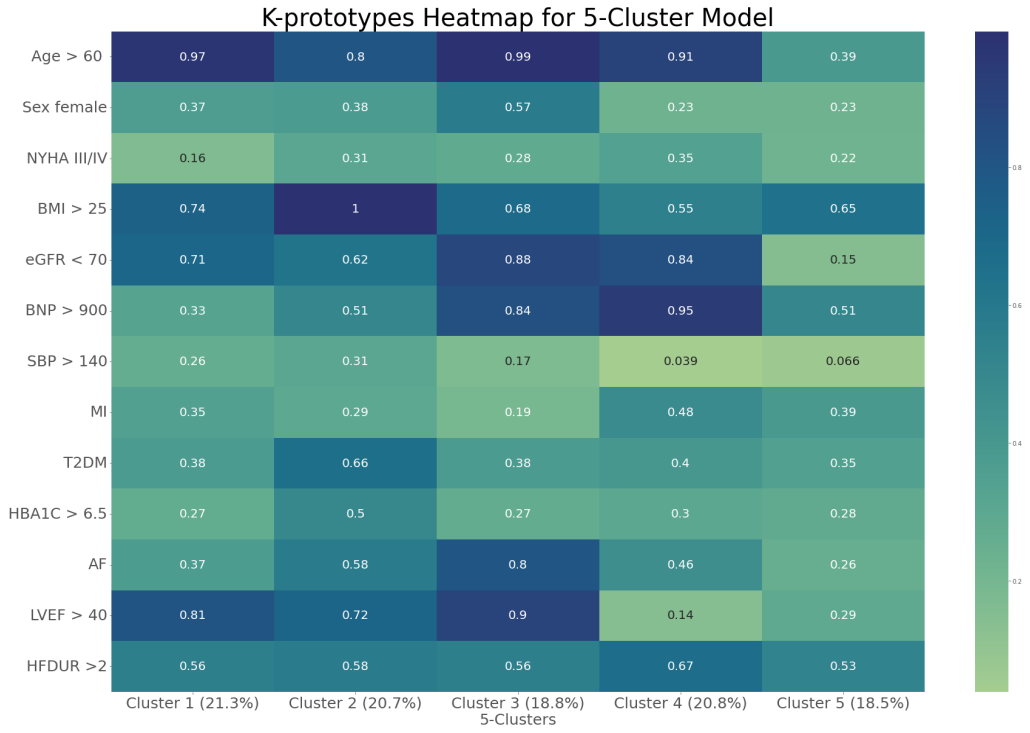

Figure A.24: Heatmap for the 4-cluster k-Prototype model

The last part of this section presents the result for the k-prototype 5-cluster model. Figure A.25, shows the covariate histograms stratified by the clusters for the 5-cluster k-Prototypes model.

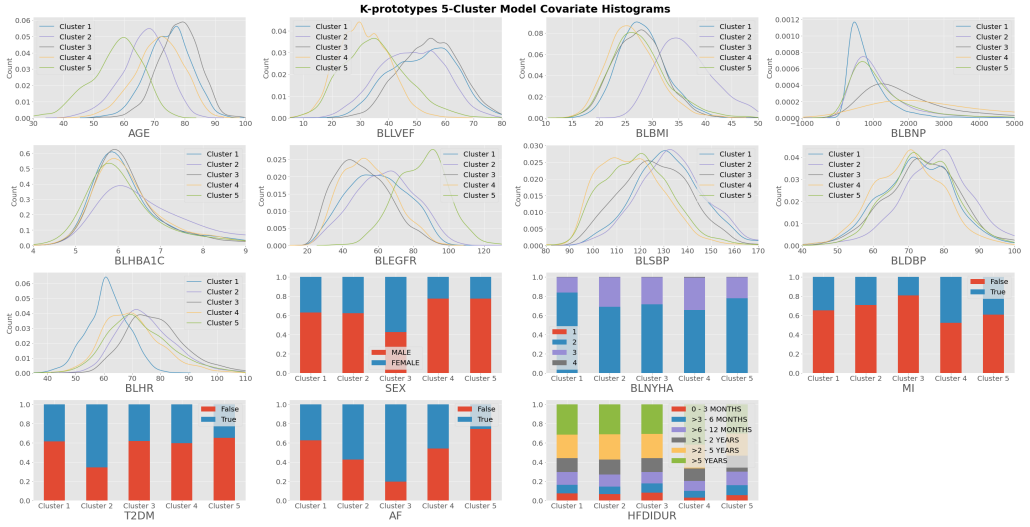

Figure A.25: Covariate histograms per cluster for the k-Prototypes 5-cluster model

In Figure A.23, the heatmap for the 5-cluster k-Prototypes model is presented.

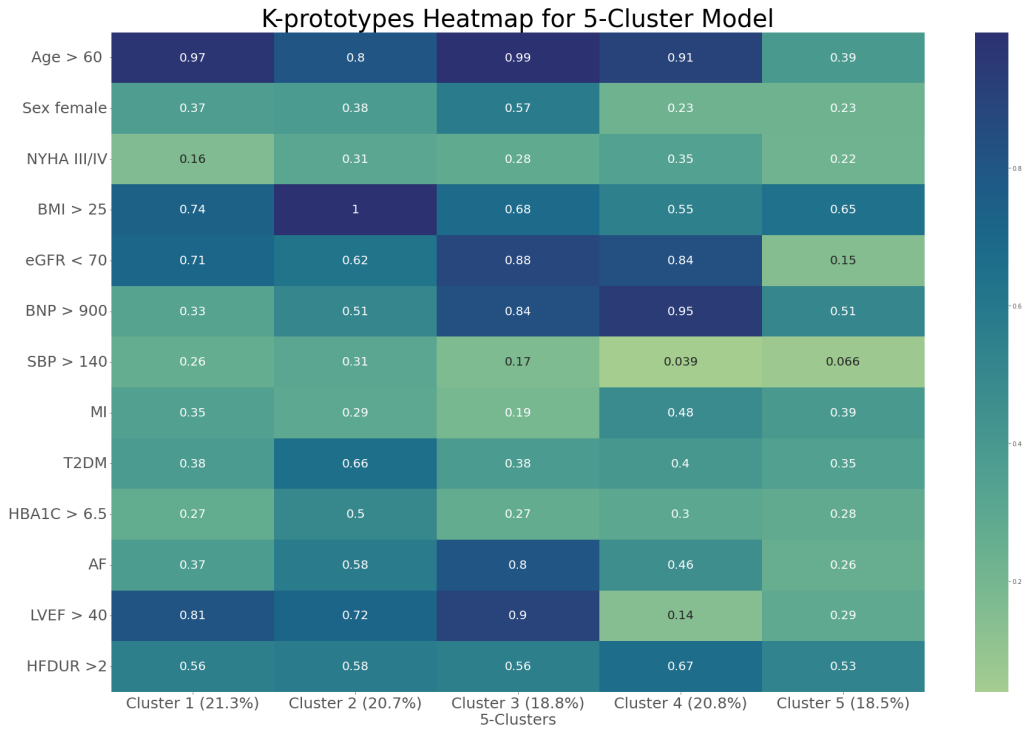

Figure A.26: Heatmap for the 5-cluster k-prototype model

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY