UNIVERSITÉ DE LIÈGE

FACULTÉ DE MÉDECINE

Département des Sciences de la Santé publique

Contribution to cluster analysis in chronic obstructive airway diseases

Halehsadat Nekoee Zahraei

Dissertation présentée
en vue de l'obtention du grade de
Docteur en Sciences de la Santé publique

2022

UNIVERSITY OF LIEGE
Faculty of Medicine
Department of Public Health

Contribution to cluster analysis in chronic obstructive airway diseases

Halehsadat Nekoee Zahraei

Supervisors

Professor R. Louis

Professor A. F. Donneau

August 2022

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my Ph.D. supervisors, Professor Renaud Louis and Professor Anne-Françoise Donneau, for their continuous support and advice throughout my time at the University of Liege. I appreciate all of the contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm they always have for my research were contagious and motivational for me, even during tough times in the Ph.D. journey. I am also special thanks for the excellent example Professor Anne-Françoise Donneau has provided as a successful woman scientist, manager, and professor.

My sincere thanks also goes to the members of the Pneumology department and Professor Renaud Louis for organizing wonderful events, walking through forests, and snow, visiting several museums and chocolate factory, and tasting some delicious food.

The Biostatistics Unit has been a source of friendships as well as good advice and collaboration; Anh Nguyet Diep, Nadia Dardenne, Justine Monseur, Sophie Klenkenberg, Francine Bonvalet and Pierre-Louis Verdin. I am especially grateful for all of my friends who had funny and enjoyable moments outside of the university, Gilles Louis, Sara Gerday, Sophie Graff, and Catherine Moermans.

For this dissertation, I would like to thank my reading committee members: Dr. Bernard Vrijens, Dr. Sylvie Streel, Dr. Catherine Moermans, and Prof. Florence Schleich for their time, interest, and helpful comments. I would also like to thank the other

two members of my defense committee, Dr. Nicolas Sauvageot, and Prof. Olivier Vandenplas, for their time and insightful questions.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. I was funded by the Federal Government Grant EOS (Excellence of science) 30565447.

My time at Liege was made enjoyable in large part due to the many Iranian friends that became a part of my life. I am grateful for the time spent with them, for keeping Iranian ceremonies alive, and also for our memorable trips.

Last but not the least, I owe my deepest gratitude to my wonderful family: my encouraging parents who raised me with a love of science and supported me in all my pursuits and dreams, my lovely brother, and my amazing and incredible sister, for supporting me spiritually throughout my life and cheering me on, even from many thousands of miles away.

Halehsadat Nekoee Zahraei
University of Liege

Liege, 28 July
2022

List of Publications

First Author

Nekoee, H., Graulich, E., Schleich, F., Guissard, F., Paulus, V., Henket, M., Donneau, A. F., & Louis, R. (2020). Are type-2 biomarkers of any help in asthma diagnosis? ERJ Open Research, 6(2), 00169–02020. https://doi.org/10.1183/23120541.00169-2020.

Nekoee, H., Guissard, F., Paulus, V., Henket, M., Donneau, A.-F., & Louis, R. (2020). Comprehensive Cluster Analysis for COPD Including Systemic and Airway Inflammatory Markers. COPD: Journal of Chronic Obstructive Pulmonary Disease, 17(6), 672–683. https://doi.org/10.1080/15412555.2020.1833853.

Nekoee, H., Louis, R., Donneau, A.F., Using multiple imputation for cluster analysis with large incomplete data. (manuscript in under-review)

Nekoee, H., Schleich, F., Guissard, F., Paulus, V., Henket, M., Donneau, A.F., & Louis, R., Clustering on the non-eosinophilic asthmatic patients. (manuscript in under-review)

Nekoee, H., Schleich, F., Guissard, F., Paulus, V., Henket, M., Donneau, A.F., & Louis, R., Clustering on the eosinophilic asthmatic patients. (manuscript in under-review)

Co-Author

Bougard, N., Nekoee, H., Schleich, F., Guissard, F., Paulus, V., Donneau, A., & Louis, R. (2020). Assessment of diagnostic accuracy of lung function indices and FeNO for a positive methacholine challenge. Biochemical Pharmacology, 179, 113981. https://doi.org/10.1016/j.bcp.2020.113981.

Louis, G., Pétré, B., Schleich, F., Nekoee, H., Donneau, A., Silvestre, A., Henket, M., Paulus, V., Guissard, F., Guillaume, M., & Louis, R. (2021). Predictors of asthma–related quality of life in a large cohort of asthmatics: A cross–sectional study in a secondary care center. Clinical and Translational Allergy, 11(7). https://doi.org/10.1002/clt2.12054.

Schleich, F., Graff, S., Nekoee, H., Moermans, C., Henket, M., Sanchez, C., Paulus, V., Guissard, F., Donneau, A., & Louis, R. (2020). Real–world experience with mepolizumab: Does it deliver what it has promised? Clinical & Experimental Allergy, 50(6), 687–695. https://doi.org/10.1111/cea.13601.

Hoge, A., Labeye, M., Donneau, A.-F., Nekoee, H., Husson, E., Guillaume, M. (2022). Health literacy and its associations with understanding and perception of front-of-package nutrition labels among higher education students. International Journal of Environmental Research and Public Health. 2022; 19(14):8751. https://doi.org/10.3390/ijerph19148751.

Louis, G., Schleich, F., Guillaume, M., Kirkove, D., Nekoee, H., Donneau, A., Henket, M., Paulus, V., Guissard, F., Louis, R., & Pétré, B., Contribution of asthma symptom intensity scales in asthma diagnosis: A prospective observational study in a secondary care center. (manuscript in under-review)

Hmaied, C., Koulchitsky, S., Nekoee, H., Gladwyn-Ng, I., Donneau, A., Seutin, V., Ketamine selectively enhances α -amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid neurotransmission onto a subgroup of identified serotoninergic neurons of the rat dorsal raphe. (manuscript in under-review)

Abstract

Chronic obstructive pulmonary disease (COPD) and asthma are complex, multi-dimensional, and heterogeneous diseases, that represent an important burden for the public health expenditure in western world. In literature, patients are divided into several different groups according to the combination of clinical, biological, and physiological characteristics, and these groups are called phenotypes. Understanding the phenotype of each patient is the first step toward effective personalized management and treatment. The common statistical approach to determine the phenotypes is cluster analysis. Cluster analysis is a well-known unsupervised learning methodology that considers multiple variables in order to create coherent subsets among a large group of patients.

In this thesis, one of the most competitive and complex statistical analysis frameworks for applying cluster analysis in incomplete large datasets was introduced. In this framework, in addition to handling the missing values by multiple imputation, the dimensions of variables were reduced, and after performing the clustering method, the final result of clustering was achieved using a novel and efficient mixture multivariate multinomial model (4M) method. The efficiency of the proposed framework was evaluated and compared using several scenarios on simulated datasets with different competitive methods for each step.

The new framework was applied to three novel specific populations of COPD and asthma. The first study was conducted on 178 stable COPD patients with the ratio of forced expiratory volume in one second (FEV1) to forced vital capacity (FVC) post

bronchodilation less than 70%, age above 40 years, and smoking history of at least 20 pack years and no clinical history of asthma before the age of 40 years. As a result, three different clusters were found, which shared similar smoking history. Including markers of systemic and airway inflammation and atopy and applying a comprehensive cluster analysis, we provide here evidence for 3 clusters markedly shaped by sex, airway obstruction, and neutrophilic inflammation but not by symptoms and T2 biomarkers.

In the next study, 426 eosinophilic patients which were defined by a sputum eosinophil count ≥ 3% were considered. On the whole cohort, cluster analysis revealed two groups identified as cluster 1 (n=276) and cluster 2 (n=150) with cluster 1 being highly atopic with achievable control of the disease with ICS in most of the cases whereas cluster 2 featured a more aggressive disease, largely non-atopic with mixed granulocytic inflammation often resisting to ICS or oral Corticosteroids (OCS).

Finally, the framework was applied to a large group of asthmatics (n=588) who were non-eosinophilic (sputum eosinophils <3%). The analysis of the whole cohort revealed two groups identified as cluster 1 (n=417) and cluster 2 (n=171) with cluster 1 displaying a low treatment burden and proportion of atopy, a neutrophilic airway inflammation, a frequent smoking history with preserved lung function but poor asthma control and quality of life while the cluster 2 essentially featured atopic patients with paucigranulocytic and partly controlled asthma.

In conclusion, our proposed framework has an effective performance compared to competing methods based on the designed scenarios on these simulated datasets. By including airway inflammatory parameters among the variables, we have provided original data on cohorts of COPD and eosinophilic and non-eosinophilic asthmatics, which indicate substantial heterogeneity between clusters and, in asthma, in particular, great differences inside each airway inflammatory phenotype. Our findings should be confirmed in multicentric studies and their clinical value assessed on longitudinal studies looking at mortality and hospitalization in COPD and exacerbation rate and lung function decline in asthmatics.

Contents

Glossary

4M mixture multivariate multinomial model

ACQ Asthma Control Questionnaire

ACT Asthma Control Test

AD Average Distance

ADM Average Distance Between Means

ARI Adjusted Rand Index

APN Average Proportion of Non-Overlap

AQLQ Asthma Quality of Life Questionnaire

BMI Body Mass Index

CAT COPD Assessment Test

CH Calinski-Harabasz index

COPD Chronic Obstructive Pulmonary Disease

CRP C-Reactive Protein

DHEAS Dehydroepiandrosterone Sulfate

DLCO Diffusing Capacity for Carbon Monoxide

FAMD Factor Analysis of Mixed Data

FENO Fractional Exhaled Nitric Oxide

FEV1 Forced Expiratory Volume in one second

FRC Functional residual capacity

FOM Figure of Merit

FVC Forced Vital Capacity

GERD Gastroesophageal Reflux Disease

HCA Hierarchical Cluster Analysis

ICS Inhaled Corticosteroids

IgE Immunoglobulin E

LABA Long Acting B2 Agonist

LAMA Long Acting Muscarinic Antagonist

LTRA Leukotriene Receptor Antagonist

MAR Missing At Random

MCA Multiple Correspondence Analysis

MCAR Missing Completely At Random

MDS Multidimensional Scaling

MNAR Missing Not At Random

OCS Oral Corticosteroids

PCA Principal Component Analysis

RI Rand Index

RV Residual Volume

SARP Severe Asthma Research Program

SC Silhouette Coefficient

SMART Single Maintenance And Reliever Therapy

SSB Sum of Square Between Clusters

SSW Sum of Squares Within Clusters

TLC Total Lung Capacity

General introduction

In heterogeneous diseases, like asthma and chronic obstructive pulmonary disease (COPD), it is crucial to describe and classify patients according to their clinical features. The sub-homogenous clusters are known as phenotypes. Cluster analysis is a statistical method that allows to classify data objects according to the information available and variables that describe those objects and their relationships. As a general rule, it is desired that objects within a cluster are similar (or related) to one another, and objects are distinct from (or unrelated to) those within other clusters. In cluster analysis, the most efficient results can be found when the highest degree of similarity (or homogeneity) is detected within the clusters and the greatest dissimilarity between them. Application of cluster analysis depends on the objects, variables, and methodology. Therefore, in contrast to other methods of defining phenotypes, cluster analysis is more data-driven; consequently, the results derived with cluster analysis may be less exposed to a priori assumptions and documented bias. When objects are assigned to clusters, these clusters must have a meaningful interpretation, and an appropriate name should be attributed according to clusters'characteristics. Within the medical field, the names of the clusters should be assigned to reflect the underlying cause of the disease as well as the clinical, physiological, and immunological features and response to treatment.

The application of cluster analysis for the classification of a population involves several major considerations. The first consideration deals with the selection of the type of population that will be studied by the cluster analysis. It is imperative to carefully select the population and individuals to analyze. In a homogeneous population

with individuals who are too similar, deriving clusters may lead to misleading results. For example, if individuals with airflow obstruction are selected as the study population and that those individuals are resistant to treatments, cluster analysis may simply reveal phenotypes related to referral patterns; in this case, insufficient response to inhaled corticosteroids. On the other hand, when the population is too disparate and individuals have a wide range of disease conditions, it may result in unexpected and ambiguous clusters or very small clusters that will not be representative of a meaningful underlying disease. Although random sampling can overcome part of these effects, this thesis stands out primarily because of its comprehensive, accurate, and innovative population selection of COPD and asthma, which leads to reliable clustering results.

The second consideration deals with the problem of missing values. In clinical research and when working with real-world datasets, missing values are pervasive and unavoidable. This issue is a significant challenge in statistical analysis. There may be several reasons for missing values, such as failure to record a specific test, miss of some questions on self-assessment questionnaires on purpose or by accident, or particularly, patients who are not able to pass some of the tests or samples, such as low-quality sputum. In terms of statistics, as there is no unique solution to deal with missing values, working on incomplete datasets is particularly troublesome. Furthermore, missing values creates serious problems when researchers are faced with insufficient statistical methods not designed to handle incomplete datasets. Most statistical softwares are usually designed with default settings that exclude missing values from the analysis and just notify a warning for the incomplete dataset. However, excluding missing values from a dataset may result in decreased power, high standard error values, wide confidence intervals, decreased precision, more bias, and reduced efficiency. Several alternative methods are available for imputation of missing values. The two main categories include simple and advanced imputation techniques. The simple approaches consist of listwise deletion, available case analysis, single imputation, the indicator method, and weighting while, advanced analysis approaches include likelihood-based methods, posterior-based approaches, and multiple imputation. Despite no discussion of the validity of these methods, cluster analysis can be applied directly to most of them. Multiple imputation has become a popular and very flexible method that takes into account the uncertainty of missing data. However, due to its complexity in combining the results to gain final clustering output and many analytical decisions, it has only been considered in a limited number of studies.

In the third consideration, researchers need to think about the variables involved in cluster analysis. The presence of mechanisms and clinical characteristics in different phenotypes will be reflected through these variables. Variable selection techniques, such as selecting variables based on researchers'opinions or selecting variables using criteria indices, are popular in clinical research. With this kind of technics, variables that measure the same thing should be avoided, since the extra noise may lead to unclear clusters. During variable selection, researchers should consider that the values of selected variables may be influenced by treatments (e.g. inhaled corticosteroids modify those associated with variable airflow obstruction) or when a disease process changes the values of variables defining the disease (e.g. variable airflow obstruction caused by inflammation may lead to irreversible obstruction due to remodeling). It is obvious, however, that different researchers would select different variables. However, there is no need for the researchers to select part of the recorded variables. Indeed, variable reduction methods exist that allow considering the whole available variables. Among those methods, the principal component analysis (PCA) reduces the number of available variables by creating new components without excluding part of the original variables set. These new components could then be analyzed more easily via cluster analysis. However, the key point in clustering is to assign all of the objects to clusters and investigate the classifications, and attribute names to clusters. Therefore, it would be necessary at this stage to refer back to the original variables.

Once the population has been defined and the methods for addressing missing values and variable reduction have been established, the dataset is ready for classification. A fourth consideration deals with methods of cluster analysis. Indeed, cluster analysis is not similar to common statistical methods in which statistical hypotheses are investigated. Cluster analysis searches for the presence of a structure in the dataset. In practice, there are numerous ways to carry out cluster analysis. Methods are simply composed of two broad categories: hierarchical and nonhierarchical. Hierarchical methods repeatedly merge (agglomeration) or divide (distributive) clusters, using distance criteria, until each object is classified. Depending on the type of variables, different approaches can be used to measure this distance. Nonhierarchical cluster analysis, like K-means, aims to find a classification of the objects which maximizes (or minimizes) criterion. In another clustering method, a mixture of multivariate normal distributions is assumed with some assumptions about the shape of clusters using a variance-covariance matrix. Based on information criteria, the optimal number of clusters can be determined, and then the objects are classi-

fied. In all of these methods, determining the optimal number of clusters is a huge challenge, and there are many different approaches to determine it. Consequently, the clustering results are highly sensitive and dependent on the method applied.

In spite of all the explained challenges, within the medical framework, cluster analyses may help clinicians understand the true patterns of COPD and asthmatic patients. This knowledge may be used to develop different pharmacological treatments and other interventions for specific phenotypic groups. In other words, when using patients'classification into several homogeneous clusters and naming them based on similar characteristics and different phenotypes, clinicians can implement more personalized treatment. Therefore, the aim of this thesis will be to apply appropriate clustering methods to specific populations.

In terms of the statistical population, we selected two datasets of the most heterogeneous diseases. The first dataset was COPD patients. Many studies have been carried out on this population via cluster analysis. However, having explained the importance of the statistical population, in this thesis, we looked out on specific population, COPD patients with FEV1/FVC ratio post bronchodilation less than 70%, age above 40 years, and smoking history of at least 20 pack-years and comprehensive parameters such as T2 biomarkers and treatments, which was an innovation. While several kinds of research have been performed on clustering in asthmatic patients, the heterogeneities inside the two the well-known asthma phenotypes, i.e., eosinophilic, and non-eosinophilic phenotypes had not been investigated so far. Therefore, an objective of this thesis will be to perform cluster analysis in patients previously segregated based on sputum eosinophilia. In addition, due to the significant effect of treatment by ICS on sputum eosinophils and other functional and clinical indices, steroid naïve and high dose ICS treated groups will also be examined separately in both eosinophilic and non-eosinophilic patients.

Since cluster analysis is a data-driven methodology, it is extremely important to carefully examine all available recorded variables. However, recent publications have concentrated cluster analysis on specific, selected sets of features and variables and then generalized it to whole parameters that have a substantial impact on the results. Therefore, in this thesis, a wide variety of new features and variables have been included, and cluster analysis has been applied to remarkable variables that, despite their influence, have not been previously investigated.

As explained, cluster analysis and handling missing values are two well-known methods. However, the combination of these two methods is a new challenge with a lot of analytical decisions. The first statistical contribution of this thesis is related to the treatment of missing values with the application of multiple imputation within the framework of cluster analysis on a large number of variables. According to the process of multiple imputation, a consensus clustering method is required in the last step to combine results derived from cluster analysis on each imputed dataset. In this thesis, a new method based on mixture multivariate multinomial distribution is proposed to solve the problem of obtaining the final clustering result.

The next statistical contribution of this thesis is to present a thorough review of the various and common methods for dealing with missing values, cluster analysis, the techniques for determining the number of clusters, dimension reduction, and finally consensus clustering. In this thesis, the efficient methods of these concepts are combined and a new framework is proposed. The efficiency of the proposed framework was assessed and compared with alternative existing methods through large comprehensive simulations.

Therefore, this thesis will explain all of these processes in seven chapters. In Chapter 1, a general definition of both COPD and asthma is presented. The phenotypes of each disease are then discussed with a comprehensive review of published cluster analysis literature on both subjects. Three different datasets used in this thesis are then presented with their related difficulties in applying cluster analysis. These datasets contain patients who suffer from COPD; both eosinophilic and non-eosinophilic asthmatic patients, and their naïve and high dose ICS treated subgroups. During this chapter, we describe the populations and considered variables, exclude and include conditions, and then the number and percentage of missing values for variables.

In Chapter 2, the different clustering methods are discussed. Three commonly applied clustering techniques are explained, partition clustering, hierarchical clustering, and model-based clustering. Finally, different criteria for cluster validation are presented.

The challenge of applying cluster analysis to a large incomplete dataset is outlined in Chapter 3. The following sections summarize the common issues of handling missing data after a presentation of the missingness pattern and mechanism. Then, the problem of dimension reduction is outlined with the review of methods

related to variable selection and variable reduction. This chapter concludes with the proposal of a consensus clustering method based on mixture multivariate multinomial distributions to solve the problem of obtaining the final clustering result.

The original statistical part of our research is presented in Chapter 4. The chapter attempts to present a new statistical framework that combines the best methods to deal with the challenges presented in Chapter 3 when applying cluster analysis to incomplete large datasets. Two commonly used frameworks that address the same issues are presented in this chapter. The performance of the proposed framework is evaluated and compared using several scenarios on simulated datasets with different competitive methods for each step. For that purpose, the percentage of the correct number of clusters, as well as the Kappa coefficient with a corresponding 95% confidence interval, is taken into account. This chapter concludes with some recommendations.

The COPD dataset and variables that make up the dataset, as well as the percent of missing values, are described in Chapter 5. In this chapter, a combination of appropriate methods is applied to classify the multidimensional incomplete COPD dataset. As a result, three distinct clusters are identified in COPD dataset.

The proposed framework is applied to eosinophilic (sputum eosinophils $\geq 3\%$) and non-eosinophilic (sputum eosinophils $< 3\%$) asthmatics patients, in Chapter 6 and Chapter 7, respectively. Since ICS treatment could be a confounding factor and that high doses of ICS may serve to define severe asthma, cluster analysis will be redone in steroid naïve and high dose ICS treated patients for both eosinophilic and non-eosinophilic asthmatics.

Figure 1 summarizes the steps of this thesis in the order in which they generally occur in cluster analysis of incomplete datasets with a large number of variables, along with related chapters. In summary, this thesis focused on the application within the medical field of cluster methods in the presence of incomplete large datasets. There is a strong emphasis on integrating methods for handling missing values and variable reduction in the clustering process. In addition to the statistical field, the findings of this thesis have an impact on the medical field as well. From a statistical point of view, based on the results of the simulation study, we provided some recommendations on the best way to apply the cluster method to incomplete large datasets. While from a medical point of view, appropriate clustering methods were applied to iden-

tify unknown dimensions of asthma and COPD diseases.

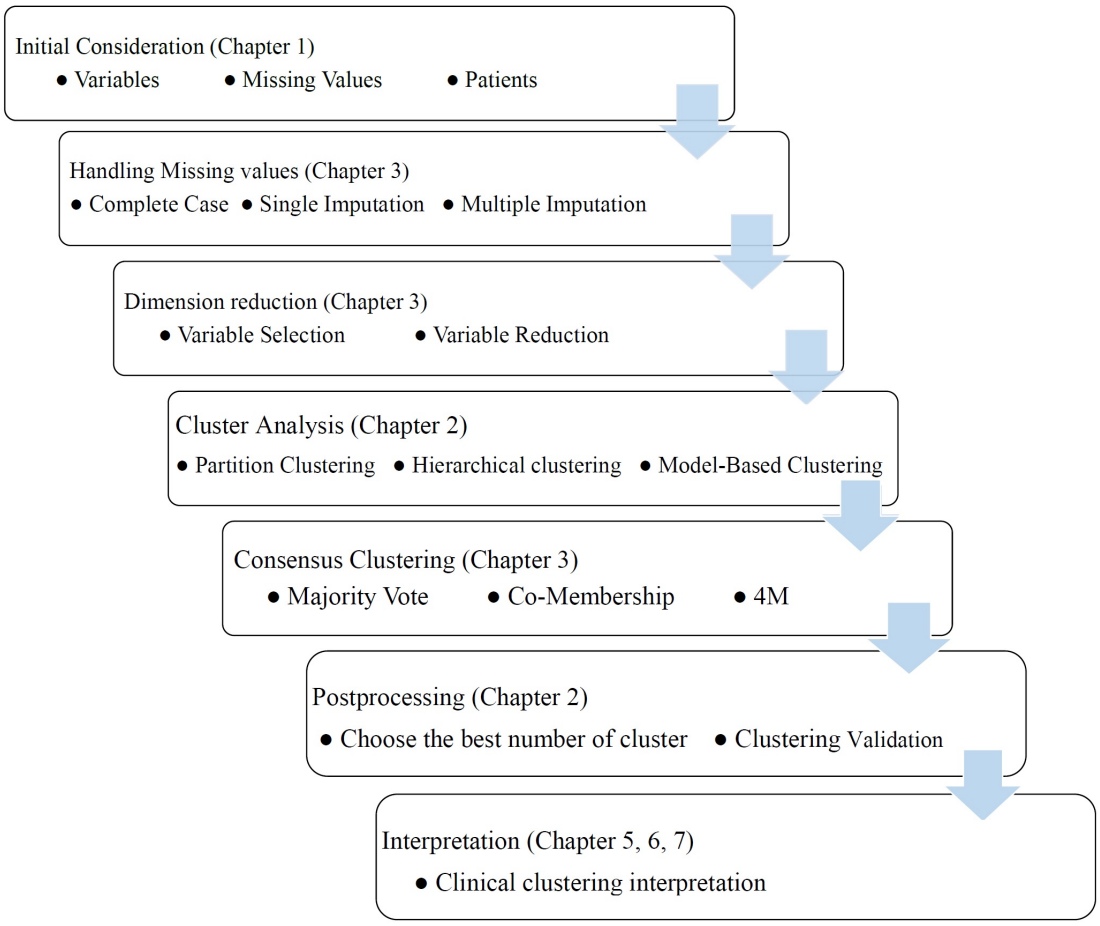

Figure 1: Schematic of the thesis (Inspired by Figure 1 of Horne et al. (2020)).

CHAPTER 1

Phenotype in COPD and asthmatic patients

COPD and asthma are widely accepted as complex and heterogeneous disorders that may have different underlying causes, and treatment response patterns, and etiologies. These two diseases can be classified into recognizable clusters, referred to as phenotypes in clinical terminology. These clusters are defined according to demographic information, clinical characteristics, lung function, and inflammation parameters measured in the patients. These phenotypes could be derived using statistical approaches such as cluster analysis. The first chapter of this thesis will focus on a literature review of the use of cluster analysis for defining COPD and asthma phenotypes in different patient populations in section 1.1. The main medical purpose of this thesis is to identify and interpret phenotypes on three datasets described in section 1.2. The first dataset will consist of data collected among patients suffering from COPD, the second dataset will focus on eosinophilic asthmatic patients, and finally, the third dataset will investigate the characteristics of non-eosinophilic asthmatic patients. The statistical challenges associated with performing cluster analysis on these datasets will then be briefly exposed in section 1.3.

1.1 Asthma and COPD phenotype

In recent years, cluster analysis has been applied as a popular method to examine the heterogeneity of patients with COPD or asthma. There are many studies of clustering on COPD or asthmatic patients which differ in their study population, sam-

ple definition, collected parameters as well as applied statistical methods. As cluster analysis includes several steps such as variable selection, method of clustering, and determination of the number of clusters, different results can emerge from these studies. In this first section, we will review the literature on clustering in both COPD and asthmatic patients.

1.1.1 COPD phenotypes

COPD is a major chronic airway and pulmonary disease affecting more than 10% of the population above 40 years. It represents a major burden for public health as it results in a high morbidity and mortality together with high costs incurred by drugs and repeated hospitalizations (Vos and et al., 2020). COPD refers to a group of lung conditions including emphysema, chronic bronchitis, asthma, and other lung diseases that can lead to breathing difficulties due to lung inflammatory conditions, irreversible narrowing of the airways, that results in decreased airflow. COPD is frequently characterized by breathlessness, coughing with phlegm, chest infections, as well as wheezing. Functionally COPD is defined by a FEV1/FVC ratio $< 70\%$ after bronchodilation by inhaled $\beta 2$ agonist and/or anticholinergic.

A major cause of COPD is smoking, which accounts for almost 90% of cases, other causes are exposure to dust or fumes at work, and air pollution. In COPD, phenotype refers to a combination of disease attributes that distinguish patients according to their clinically important characteristics, such as gender, exacerbation, symptoms, response to treatment, and rate of disease progression (Han et al., 2010).

These phenotypes divide patients into several clusters with common features and clinically meaningful outcomes that help patients to receive effective care and achieve better clinical results. The exploration of COPD using their phenotypes has a potential impact on pharmacological and non-pharmacological management of COPD. Identifying the phenotypic approaches is the first step toward more personalized management and treatment of COPD patients.

The phenotypes of COPD can be investigated in several aspects. Snider (1989) was the first to introduce and define the different types of phenotypes with a specific statistical method, known as non-proportional Venn diagram. Chronic bronchitis, emphysema, and asthma were the three subgroups described by the authors, with subgroups overlapping with each other.

Among all phenotypes introduced and described in the literature, we considered studies utilizing cluster analysis on multidimensional COPD datasets, and the results were clear and understandable. Miravitlles et al. (2013) found three different phenotypes when considering the different responses to available treatments. These three distinguished phenotypes were the exacerbator, the overlap COPD-asthma, and the emphysema-hyperinflation. Patients with the exacerbator phenotype presented at least two exacerbations in the previous year, and in addition to long-acting bronchodilators, anti-inflammatory medications may be necessary.

The overlap COPD-asthma phenotype was characterized by an increase in airflow variability and incompletely reversible obstruction of airflow. This phenotype presented a good therapeutic response to inhaled corticosteroids, as well as bronchodilators, which may be related to the underlying inflammation profile. Moreover, the emphysema phenotype did not respond well to the currently available anti-inflammatory drugs, and long-acting bronchodilators in conjunction with respiratory rehabilitation were the treatment of choice.

The same year, Spain's COPD guidelines (GesEPOC (2012); Miravitlles et al. (2013)) were the first to introduce the phenotype into clinical practice (Miravitlles et al., 2013). Then, detailed analysis of national guidelines (across Europe and Russia; Miravitlles et al. (2016)) showed a high variability in detecting COPD phenotypes, including classic COPD phenotypes of chronic bronchitis and emphysema (Miravitlles et al., 2016).

In 2019, a new definition of phenotype was introduced in the GOLD guidelines management of stable COPD. According to these guidelines, the initial treatment was based solely on phenotypes of four groups (A, B, C, D). Dyspnoea and exacerbations were proposed as two main groups with individualized treatment algorithms. In addition, the blood eosinophil count has been proposed as a biomarker for predicting the response to treatment with inhaled corticosteroids (Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (Gold et al., 2019)).

The GOLD guidelines were published to help clinicians diagnose and treat COPD. There is widespread agreement in all guidelines that COPD is an extremely complex disease with a high mortality rate. It is still a challenge to identify accurate subgroups

or phenotypes, and introduced phenotypes have progressed through time. Using the latest technology and medical knowledge, current research highlights additional clinical parameters such as biomarkers, which are becoming the basis for clinical phenotypes.

Besides the current GOLD criteria, unsupervised COPD clustering studies have brought insights into the importance of co-morbidities and systemic inflammation as components accounting for disease variability (Burgel et al., 2017). Indeed, over the last decade, cluster analysis has become a popular method to examine heterogeneity of patients with COPD (Pistolesi et al., 2008; Paoletti et al., 2009; Burgel et al., 2010; Garcia-Aymerich et al., 2011; Fens et al., 2013; Burgel et al., 2017).

There are studies that used demographic variables, symptoms, spirometry, imaging, and comorbidities to derive the clusters (Burgel et al., 2012). An overview of cluster analysis to identify phenotypes in COPD can be found in Table 1.1, which contains population, the considered variables, the clustering method, and the main results of cluster analysis.

However, there are not many studies using clustering that investigated the airway inflammatory component and the atopic status in a large cohort of COPD. Since it has been recognized that some COPD patients may express T2 biomarkers (Miravitlles et al., 2013; Cataldo et al., 2017), a section of our thesis aims to determine whether the T2 trait is common and strong enough to identify a cluster in COPD patients denying any history of asthma that could have started before the age of 40 years.

Table 1.1: Overview of papers using cluster analysis to identify COPD phenotypes

Reference	n	Setting	Population	Data used to build clusters	Types of analyses	Main results
Altenburg et al. (2012)	65	Single center, tertiary care, pulmonary rehabilitation (Groningen, The Netherlands)	Moderate to very severe airflow limitation Referred for rehabilitation	Age, BMI, quadriceps force, body plethysmography, and exercise testing	K-means	2 phenotypes: (i) worse lung function and exercise capacity, worse quadriceps force, and better response to exercise training (ii) better lung function and exercise capacity and less response to exercise training
Burgel et al. (2010)	322	Multicenter cohort (Initiatives BPCO), tertiary care (France)	Mild to very severe airflow limitation Outpatients	Age, history, symptoms, spirometry, BMI, exacerbations, health, psychological status,	PCA, HCA (Ward's)	4 phenotypes: (i) young subjects with severe respiratory disease, cachexia (ii) older subjects with mild airflow limitation and mild comorbidities (iii) young subjects with moderate to severe airflow limitation, but few comorbidities (iv) older subjects with moderate to severe airflow limitation and high rates of cardiovascular comorbidities
Burgel et al. (2012)	527	Single center, tertiary care (Leuven, Belgium)	Mild to very severe airflow limitation Outpatients and COPD patients identified as part of a lung cancer screening study	Age, history symptoms, health status, body plethysmography, DLCO, CT-scan, physician-diagnosed comorbidities	PCA, MCA, HCA (Ward's)	3 phenotypes: (i) younger patients with severe respiratory disease, cachexia, and low rates of cardiovascular comorbidities. (ii) older patients with less severe airflow limitation, but often obese and with high rates of cardiovascular comorbidities and diabetes. (iii) mild to moderate airflow limitation, absent or mild emphysema, absent or mild dyspnoea, normal nutritional status, and limited comorbidities
Fens et al. (2013)	157	Population-based survey (Utrecht,The Netherlands)	Mild to moderate airflow limitation COPD patients identified as part of a lung cancer screening study	History, symptoms, health status, comorbidities, spirometry, DLCO, CT-scan, breathomics (electronic nose)	PCA, HCA (Ward's), K-means	4 possible phenotypes: (i) mild COPD (ii) moderate airflow obstruction with chronic bronchitis and emphysema (iii) asymptomatic emphysema with preserved lung function (iv) high symptoms, preserved lung function
Garcia-Aymerich et al. (2011)	342	Multicenter study, tertiary care (Spain)	Mild to very severe airflow COPD patients recruited after a 1^{st} hospitalization	History symptoms, health status, body composition, plethysmography, CT-scan, biology (sputum and serum), and exercise testing	K-means	3 phenotypes: (i) severe respiratory COPD (ii) moderate respiratory COPD (iii) systemic COPD (high rates of cardiovascular comorbidities)
Paoletti et al. (2009)	415	Single center, tertiary care (Florence, Italy)	Mild to very severe airflow limitation Outpatients	History and symptoms, body plethysmography, DLCO, and chest X-ray	MDS, PCA, MCA, K-means	2 phenotypes: (i) predominant airflow obstruction (ii) predominant parenchymal destruction
Pistolesi et al. (2008)	322	Single center, tertiary care (Florence, Italy)	Mild to very severe airflow limitation Outpatients	13 comorbidities	SOM, HCA (Ward's)	5 possible comorbid phenotypes: (i) less comorbidity (ii) cardiovascular (iii) cachectic (iv) metabolic (v) psychological with no difference in systemic inflammation

1.1.2 Asthma phenotypes

Asthma is a chronic airway disease affecting more than 300 million people worldwide and 5 – 10% of the population in the western world (Vos and et al., 2020). The disease represents a major cost for public health authorities largely due to drug costs (Nunes et al., 2017). However, the number of patients with poorly controlled asthma remains substantial in spite of improved asthma pharmacological treatment (Pavord et al., 2018).

The Global Initiative for Asthma (GINA) describes asthma as a heterogeneous disease, usually characterized by chronic airway inflammation, defined by a history of respiratory symptoms such as wheeze, shortness of breath, chest tightness, and cough that vary over time and intensity, together with variable expiratory airflow limitation (GINA, Reddel, et al., 2021). There are many factors involved in this, one of them being asthma heterogeneity (Gold et al., 2012; Wenzel, 2016).

Asthma is characterized by heterogeneity evident in varying exacerbation risks and treatment responses. The classification of asthma into phenotypes was made in order to maximize management for different patient severity levels (Wenzel, 2012). Since the classification of asthma is a complex process based on the multidimensional nature of the disease; such as symptoms, lung function, systemic and airway inflammation and treatment. Therefore, cluster analysis can provide an effective way of identifying asthma clusters.

The first study that generated interest in clustering methodology within asthma field was conducted by Haldar et al. (2008) in Leicester, UK. As this article presents a cluster analysis study in the field of asthma with clear and understandable results, the findings of this study will be explained in more detail. A two-step Ward's hierarchical and subsequent K-means cluster analysis were performed in three independent asthma populations (refractory asthma population managed in secondary care, primary care with predominantly mild to moderate disease, and refractory asthmatics from clinical trials).

After a subjective selection of 16 clinically relevant variables (PEF Variability, SPT Cat fur, SPT Dog dander, SPT D. Pteronyssinus, SPT Grass Pollen, Nocturnal Symptoms, Daytime Symptoms, Activity Symptoms, Dyspnoea, Wheeze, Anxiety Score, Depression Score, Exhaled NO 50ml/sec, Sputum Eosinophils, Blood Eosinophils,

FEV1 response to BD), Principal Component Analysis (PCA) was conducted to reduce the 16 variables into five components. According to the results of cluster analysis, primary care patients were divided into three clusters, while secondary care patients contained four clusters. In these two populations, two common clusters were identified: the first cluster was composed of early-onset atopic asthmatics whereas the second featured obese, often female, non-eosinophilic asthmatics.

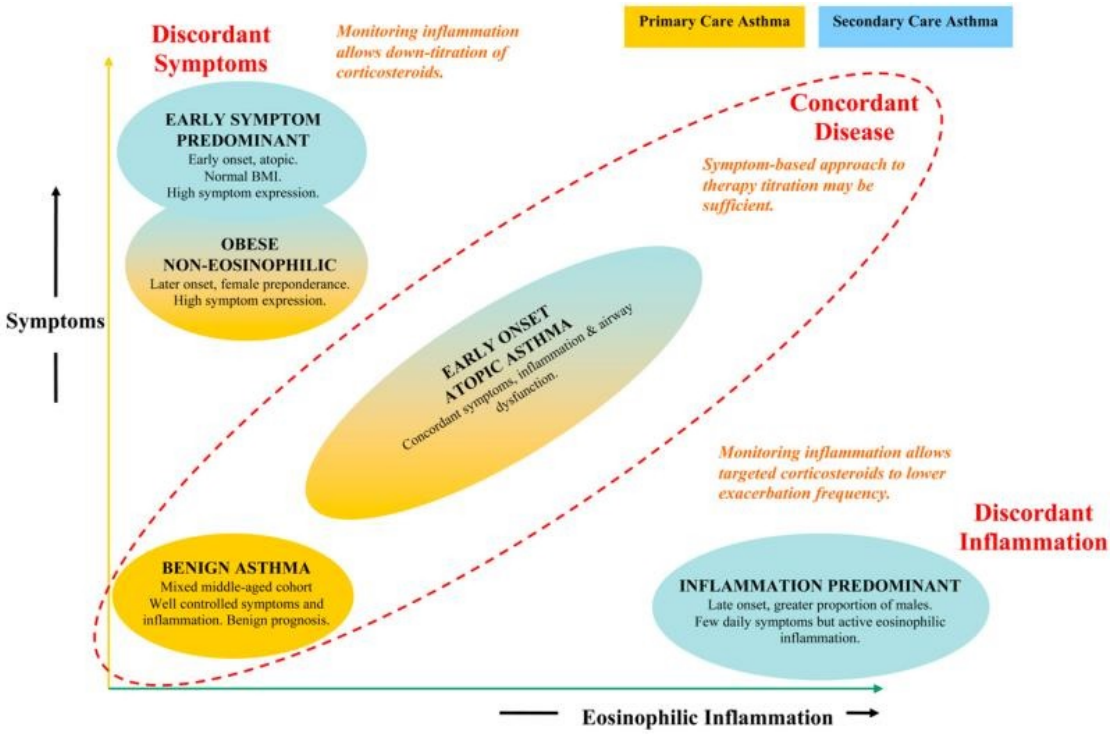

Figure 1.1: Summary of asthma phenotypes identified by cluster analysis in the primary- and secondary-care populations (Haldar et al., 2008).

In the primary care, there was a third 'benign asthma' cluster, while two additional cluster were found in the secondary care. One featured early onset symptom predominant non-eosinophilic asthmatics whereas the other featured late onset, often male, eosinophilic asthmatics with few symptoms (Haldar et al., 2008). Figure 1.1 summarizes the asthma phenotypes identified using cluster analysis in primary and secondary care populations, which was first proposed by Haldar et al. (2008).

Consequently, cluster analysis identifies phenotypes that respond differently to treatment and may require different treatment strategy. Indeed, the SMART (single maintenance and reliever therapy) strategy grading the dose of ICS based on symptom expression (Rabe, 2004) might not be suitable in those showing symptoms with-

out airway eosinophilic inflammation and those with intense eosinophilic inflammation and few symptoms alike.

The evaluation of sputum inflammatory cells plays a key role in phenotyping and clustering asthmatic patients. In Simpson et al. (2006), sputum cells were evaluated in order to differentiate between inflammatory subtypes of asthma, and then clinical features were compared across the subtypes of asthmatic patients. Sputum eosinophil and neutrophil proportions were used to cluster patients based on the 95^{th} percentile of a healthy control group.

The authors identified the following four clusters of asthma: (1) eosinophilic (sputum eosinophil cutoff, > 1.01%), (2) neutrophilic (sputum neutrophil cutoff, > 61%), (3) mixed granulocytic (high eosinophil and neutrophil counts), and (4) paucigranulocytic (low eosinophil and neutrophil counts). Subsequently, Hastie et al. (2010), and Schleich et al. (2013) have extended analyses based on this classification. In terms of clinical characteristics, they found that the mixed granulocytic phenotype showed the most impaired lung function while the eosinophilic phenotype was more prone to exacerbation.

The neutrophil-predominant subtype had a significantly later age of asthma onset (Kaur and Chupp, 2019). A recent longitudinal study from the SARP (severe asthma research program) has shown that mixed granulocytic phenotype was associated with accelerated lung function decline (Hastie et al., 2021), which supports the association between poor lung function and the mixed granulocytic phenotype found in cross-sectional studies. All these observations highlight the value of sputum analysis in the interpretation of clinical asthma outcomes and therefore the importance of including sputum variables in a comprehensive cluster analysis of asthmatics (Louis and Schleich, 2021).

There are many other asthma studies where researchers separated patients into several phenotypes based on symptoms, triggers, age at onset of disease, underlying inflammation and other routine parameters (Moore et al., 2010; Kuhlen et al., 2014; Newby et al., 2014; Denton et al., 2021). An overview of cluster analysis to identify phenotypes in asthma can be found in Table 1.2, which contains population, the considered variables, the clustering method, and the main results of cluster analysis.

To the best of our knowledge there has been no cluster analysis specifically focus-

ing on eosinophilic asthma on the one hand and non-eosinophilic asthma on the other hand. The clustering of eosinophilic asthma might be of particular interest as a study has suggested that there might be heterogeneity in the function of eosinophils in human (Mesnil et al., 2016).

When evaluating inflammatory asthma phenotypes, use of inhaled corticosteroids (ICS), which is the mainstay of asthma treatment, may be a confounding factor. Numerous studies in adults have shown that initiating or increasing ICS treatment is associated with a significant decrease in sputum eosinophil. In contrast, discontinuing ICS results in an increase in eosinophils (Hancox et al., 2012).

Use of ICS also improves several asthma outcomes like lung function and asthma control while reducing exacerbation. Therefore, in this thesis, we not only performed cluster analysis on the whole cohort of selected asthma population but also on subgroups divided into ICS naïve patients and those receiving high dose ICS. A supplementary reason for isolating the high dose ICS treated patients is that the patients from this group may be considered as severe asthmatics, a group of patients which deserves further research (Chung et al., 2014). In Section 1.2.2, the dataset for eosinophilic asthmatic patients will be presented, and in Section 1.2.3, it will be discussed for non-eosinophilic asthmatic patients.

Table 1.2: Overview of papers using cluster analysis to identify Asthma phenotypes

Reference	n	Setting	Population	Data used to build clusters	Types of analyses	Main results
Haldar et al. (2008)	1. 184 2. 187 3. 68	Single center, tertiary care (Leicester, UK)	1.Primary care with predominantly mild to moderate disease 2.Refractory asthma population managed in secondary care 3.Refractory asthmatics from clinical trials(longitudinal study)	16 clinically variables	PCA, HCA (Ward's), K-means	1. 3 phenotypes (i) Early-Onset Atopic Asthma (ii) Obese Non-eosinophilic (iii) Benign Asthma 2. 4 phenotypes (i) Early Onset, Atopic (ii) Obese, Non-eosinophilic (iii) Early Symptom Predominant (iv) Inflammation Predominant 3. 3 phenotypes (i) Obese female (ii) Inflammation predominant (iii) Early symptom predominant
Moore et al. (2010)	726	Multi center (SARP), (US)	Persistent asthma who have undergone detailed phenotypic characterization	34 core variables	Selected variables, HCA (Ward's)	5 phenotypes (i) Early onset atopic asthma with normal lung function treated with two or fewer controller medications and minimal health care utilization (ii) Early-onset atopic asthma and preserved lung function (iii) Mostly older obese women with late-onset non-atopic asthma (iv, v) Clusters 4 and 5 have severe airflow obstruction with bronchodilator responsiveness but differ in to their ability to attain normal lung function, age of asthma onset, atopic status, and use of oral corticosteroids
Kuhlen et al. (2014)	139	Single center, tertiary care (South Carolina, US)	Mild, moderate, and severe persistent asthma	17 demographic, clinical variables	K-means	5 phenotypes (i ii) Early-onset atopic asthma and reduced lung function but differed in medication requirement and health care utilization (iii) Older obese women with late-onset asthma, less atopy, and mildly reduced forced expiratory volume over first second (iv, v) Atopic asthma with severe obstruction but differed in bronchodilator response, age of onset, and oral corticosteroid use
Newby et al. (2014)	349	Multi centers, tertiary care (Belfast, Leicester, London , Manchester, UK)	Severe refractory asthma	23 variables	Multiple imputation , PCA, cluster mixture analysis model	5 phenotypes (i) Atopic with early onset disease (ii) Obese with late onset disease (iii) The least severe disease (iv) The eosinophilic with late onset disease (v) Significant fixed airflow obstruction
Denton et al. (2021)	1175	Multi centers, tertiary care (worldwide)	Adults with severe asthma	Immunoglobulin E, blood eosinophils, and fractional exhaled nitric oxide	HCA (Wardâs), K-means	5 phenotypes (i) Highly symptomatic, older females with elevated BMI and frequent exacerbations (ii) Older females with lower BMI and frequent exacerbations (iii) Older, highly symptomatic, lower BMI, and preserved lung function (iv) Younger, long duration of asthma, elevated BMI, and poor lung function (v) Younger males with low BMI, poor lung function, and high burden of sinonasal disease and polyposis

1.2 Illustrative datasets

This section describes the three datasets used in this thesis. The primary medical focus of this thesis is to perform cluster analysis on several populations not previously investigated. Therefore, it is essential to select the individuals who compose the population precisely. Since cluster analysis is a data-driven methodology, it is extremely important to carefully select the patients and then examine them using all available recorded variables in the datasets before determining the appropriate clustering method.

1.2.1 Chronic obstructive pulmonary disease (COPD) dataset

The first dataset included 178 stable COPD patients recruited from ambulatory care in the COPD clinic in the Pneumology Department of the University hospital of Liege. The study had a general agreement from the ethics committee to use clinical data collected from routine practice to make retrospective reports. The protocol was approved by the Hospitalo-Facultaire Universitaire ethics committee, Liege (institutional review board 2005/181). Every patient attending ambulatory clinic care signs an informed consent stating that they accept this principle. As explained in the previous section, cluster analysis is crucial in COPD patients given the heterogeneity within patients satisfying the current definition. In this dataset, cluster analysis was applied on a specific population with an emphasis on novelty.

Selection criteria to be referred to the COPD clinic were symptomatic patients (including at least one of the three following symptoms: dyspnea, cough and sputum production) with FEV1/FVC ratio post bronchodilation less than 70%, age above 40 years and smoking history of at least 20 pack years. Care was taken to exclude asthma history starting before the age of 40. At the COPD clinic, the patients had systematic pre and post bronchodilation spirometry, sputum induction, blood sampling, and completed the self-administered CAT questionnaire. From the clinical data, a comprehensive list of 84 variables was derived and categorized into six categories i.e. (1) demographics, (2) pulmonary function tests, (3) treatment features, (4) blood cell counts and systemic inflammatory markers, (5) atopic status, and (6) sputum cell counts and microbiology.

In this dataset, more than half of the patients were male (54.49%) with age ranging from 40 to 84 years and median age was 64.5. Overall the population displayed a

normal weight (median body mass index was 23.62kg/m^2). Patients had a consistent tobacco consumption history with a median pack/year of 37 and a median of smoking duration was 43 years. As a result of repeated exacerbations, 39% of patients had been treated with course of oral corticosteroids (OCS) the year prior to the investigations and 15% had experienced at least one course of OCS. Sixty-three percent of patients had at least one course of antibiotics and 12% had at least one course the year prior to the investigation. The percentage of missing values ranged from 0% to 23% and 75% of patients presented at least one missing value. Descriptive analysis is provided in Chapter 5 with additional explanations.

Atopic status was defined based on positive skin prick tests or specific IgE (> 0.35 kU/l; Phadia; Groot-Bijgaarden, Belgium) towards common aeroallergens including mites, cat and dog dander, grass and birch pollens and molds mixture. As part of the lung function test, spirometry was used (Spiro bank; MIR, Rome, Italy). Regardless of FEV1 and FEV1/FVC baseline ratio, all patients underwent a post-bronchodilator (reversibility) test.

The patients were premedicated with 400 μg of salbutamol one puff at a time into the spacer (pMDI+spacer), and sputum was obtained by inhaling hypertonic sodium chloride solution (NaCl 4.5%) in combination with salbutamol (Delvaux, 2004) using an ultrasonic nebulizer (Ultra-Neb 2000, De Vilbiss, Somerset, PA, USA) at a flow rate of 0.9mL/minute. Induction was performed when post-bronchodilation FEV1 was less than 65% of the predicted value with physiologic fluid (NaCl 0.9%) given in combination with salbutamol. For a total of 15 minutes, each patient inhaled the aerosol for three consecutive periods of 5 minutes. The safety limit was maintained by monitoring FEV1 every 5 minutes and stopping the induction after FEV1 delineated by > 20% from the post-bronchodilation value.

Sputum was collected in a plastic container, weighed, and homogenized by adding three volumes of PBS, vortexed for 30 seconds, then centrifuged at 800g for 10 minutes at 4°C. The supernatant of a cell pellet was separated and resuspended in a solution containing 5 mM DTT without Ca2+ and Mg2+, filtered, and used for a squamous and total cell count using a manual hemocytometer. Trypan blue exclusion was used to check cell viability. The differential was performed on cytospins stained with Diff-Quick after counting 500 cells (Quaedvlieg et al., 2009).

Fractional exhaled nitric oxide (FeNO), sputum cell counts, blood cell counts, and

systemic markers were used as indicators of inflammation parameters. FeNO was measured using NIOX (Aerocrine, Solna, Sweden) at a flow rate of 50ml/s before spirometry (Demarche et al., 2016). C-reactive protein (CRP), fibrinogen, blood leucocyte counts and total and specific (RAST) serum IgE were determined by routine laboratory analysis at Liege University Hospital.

1.2.2 Eosinophilic Asthmatic Patient Dataset

The second dataset included 426 eosinophilic patients, defined by a sputum eosinophil count ≥ 3% and recruited from the asthma clinic of Liege University between 2011 and 2020. The current study included comprehensive categories of parameters, such as demographics, standard routine investigations of asthma, pulmonary function tests, blood tests, atopic status, sputum, microbiology. Comorbidities such as gastroesophageal reflux disease (GERD), allergic rhinitis and nasal polyposis were identified by history taking and demonstration of objective investigation collected in the electronic medical record. Although most of the parameters were similar to what we described for COPD dataset, additional factors such as asthma quality of life questionnaire (AQLQ) and Juniper's asthma control questionnaire (ACQ) (Juniper et al., 1999), and asthma control test (ACT) (Nathan et al., 2004) were used to estimate quality of life and asthma control, respectively. AQLQ was measured over the last 14 days, ACQ over the last 7 days and ACT over the last 28 days.

There were more than half of the patients who were female (55%). Median age was 53 years and and the median age at diagnosis was 30 years. The patients displayed a slight overweight (median body mass index was $26 kg/m^2$). More than half of the patients were atopic (56%). Smoking status was categorized into three groups: never-smokers, ex-smokers (who had quit smoking at least 6 months ago), and current smokers. Forty-five percent of patients had a smoking history. The percentage of missing values ranged from 0% to 77% and 91% of patients presented at least one missing value.

The definition of an exacerbation in the year prior to the visit was a three-day course of OCS for non-OCS treated patients and a quadrupling of the dose for OSC maintenance patients. As asthma maintenance therapy relies heavily on ICS, which are well established to be effective, but can also influence the phenotype, we performed additional analyses on a steroid naïve group (n = 114) and a group a patients treated with high dose ICS (> 1000 μg/d equivalent beclomethasone) (n = 239) se-

lected from the whole cohort. As part of Chapter 6, we present a comprehensive description of eosinophilic cohort as well as the two subgroups based on ICS.

1.2.3 Non-eosinophilic Asthmatic Patient Dataset

The third dataset included 588 non-eosinophilic patients, defined by a sputum eosinophil count< 3% and recruited from the asthma clinic of Liege University between 2011 and 2020. Except, comorbidities variables such as GERD and nasal polyposis which were not available, the parameters used in this section are exactly the same as those presented for eosinophilic patients. The median age was 50 years and the median age at diagnosis of 33 years. The female gender was dominant (63%), almost half reported a smoking history (48%), while less than half were atopic patients (46%). The median body mass index (BMI) was $26 kg/m^2$ indicating a slight overweight. The percentage of missing values ranged from 0% to 66% and 89% of patients presented at least one missing value. For the same reason as mentioned above, further analyses were conducted on a group of steroid naïve patients (n = 279) and a group receiving high doses maintenance ICS (n = 135). The comprehensive descriptive analysis as well as classification results for these three cohorts based on the proposed clustering framework for incomplete datasets will present in Chapter 7.

1.3 Problem Statement and Challenges

The purpose of this section is to introduce the statistical challenges involved with cluster analysis on the three previously introduced datasets. As already mentioned, one unavoidable problem in datasets that contain many variables, is the presence of missing values. Dealing with missing values is a general challenge in real-world data analysis in epidemiological and clinical research, specifically in cluster analysis where the objective is to assign patients to clusters based on similarity.

Cluster analysis can be directly applied on majority of method of handling missing values. Among all methods of handling missing values, multiple imputation is a well-designed that accounted for the uncertainty in missing data. However, applying cluster analysis using multiple imputation is complicated. The combination of cluster analysis and multiple imputation requires an integrated framework that incorporates several steps and involves a considerable number of analytical decisions. In multiple imputation, several complete imputed datasets are generated, each of which must be analyzed separately. The principle of this method is for estimating the

parameters and the results can easily be combined to obtain the final result. However, this method is not easy to apply on cluster analysis and this makes the challenge associated with using multiple imputation in cluster analysis that we addressed this issue by using a mixture multivariate model in this thesis.

Missing values as well as dimension reduction are two important factors that are rarely considered simultaneously in cluster analysis studies. Both of these statistical challenges appear during the descriptive analysis of the datasets, and so before applying the clustering process. Therefore, decisions must be taken, and each considered decision at any step of the process will have a significant impact on the final clustering output. Each step involves several competitive methods. First, there are different ways to handle missing values, and in particular, the number of multiple imputations to consider, and how to predict missing values for each variable based on its type. Second, the number of variables that should be considered for clustering is still an open question as there is still no priority between selecting the appropriate variables or applying the principal components for variables reduction. Third, there are multiple methods to determine the number of clusters and classification methods, which will yield different results. Finally, if multiple imputations are applied, a consensus clustering method is required to merge all the clustering results and report the final and unique result. In the literature, consensus clustering and integrated framework have not been investigated in detail for clustering on incomplete datasets. As a result, the complexity of the study requires a systematic approach. An integrated framework has been proposed in this thesis. Furthermore, there are no literature references for comparing methods in the cluster analysis field. Therefore, huge comprehensive simulation scenarios are required to compare and evaluate each and combination of the methods in the framework.

1.4 Conclusion

In the literature, asthma and COPD are known as heterogeneous and complex diseases with a variety of dimensions which results in the identification of different phenotypes. The phenotypes of each disease depended on the way the population is selected and the variables that were taken into consideration. This chapter attempted to provide a general review of published literature dealing with cluster analysis to determine the phenotypes in both diseases. Three different target populations were defined in the framework of this thesis. The first studied population is composed

of adults who suffer from COPD without any previous history of asthma, the second includes patients with eosinophilic asthma and the third includes patients non-eosinophilic asthma. In order to further investigate and evaluate heterogeneity in the eosinophilic and non-eosinophilic asthmatic patients, patients were also divided into two distinct groups according to ICS treatment, that are a steroid naïve and high dose ICS treated groups. There were several criteria for including and excluding samples in this study, which are described in detail. In the general description of three datasets, the number and percentage of missing values were reported. For applying cluster analysis using multiple imputation to all datasets defined in this chapter, several statistical challenges were encountered, therefore, a detailed explanation of the difficulties of applying cluster analysis to these incomplete datasets with multi-dimensional variables was provided to close this chapter.

CHAPTER 2

Cluster Analysis

This chapter introduces the concept of cluster analysis as well as the different steps and options that underlie the application of these statistical methods. In Section 2.1, the basic concepts of clustering are introduced. First, a mathematical definition of data clustering is presented, with methodology, and types of clustering. As the process of most clustering methods involves grouping objects according to their similarity or dissimilarity using distance measures, this section also proceeds with diverse measures of two objects'distance. Three well-known clustering methods are presented in section 2.2, including partition clustering methods, hierarchical clustering, and model-based clustering. The basic issue of cluster analysis is determining the proper number of clusters. Throughout section 2.2, we attempt to describe how to determine the appropriate number of clusters for each clustering method. This chapter concludes by reviewing the validity criteria which assess the quality of clustering results. In Section 2.3, three validity statistical methods are illustrated: internal, external, and cluster stability.

2.1 Basic concepts

In the early stages of data analysis, when little knowledge of the data is available, cluster analysis is often used to gain an understanding of the similarity or dissimilarity between objects. Using this approach, objects are grouped into clusters with respect to variables, characteristics, or attributes of interest such that objects within

the same cluster are more like each other than those in other clusters. The objective that discriminates objects into homogeneous and distinct groups generally seems to be obvious, however, its formulation is not so clear in theory. First, a formal structure for clustering must be defined, and this is explained in Section 2.1.1. Furthermore, the first definition of clustering refers to objects that are near or far from one another and describes it as similarity or dissimilarity between objects. Therefore, Section 2.1.2 provides different options of measuring similarity in data and presents some special properties of these measures.

2.1.1 Notation

Let's consider a set of N objects, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T$, where each object is recorded for p-dimensional variables. The observed value of the j^{th} variable of the object i is denoted by, $x_{ij}, i = 1, \cdots, N; j = 1, \cdots, p$. Therefore, the p-dimensional vector of $\mathbf{x}_i = [x_{i1}, \cdots, x_{ip}]$ presents the recorded data for i^{th} object over p variables. The aim of cluster analysis is to partition the object $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T$ into $k \in \{2, 3, \cdots, n-1\}$ clusters denoted $\{C_1, C_2, ..., C_k\}$ where C_k is the result of the partitioning of N objects into k clusters. Hard or crisp clustering is a method that involves objects belonging to only one cluster and follows the following requirements:

1. $C_1 \cup ... \cup C_k = \mathbf{X}$, each object must belong to one of the clusters,
2. $C_i \neq \varnothing$, all clusters must contain at least one object,
3. $C_i \cap C_j = \varnothing$ for $i \neq j$, a single object cannot be part of more than one cluster at the same time.

In addition, there are two special clustering cases, when $k = 1$ where all objects are gathered into one single cluster and when each object makes its own cluster, $k = n$. Those situations should be avoided, because not of interest.

When clustering is complete, each object is assigned to one cluster, $\mathbf{x_i} \in C_k$. In this case, a vector of clustering labels can be defined for all objects. Hence, this clustering labels'vector is an integer vector $(c_1, c_2, ..., c_N)$ with values between 1 and k, where c_i is the clustering label for i^{th} object.

As an example, let's consider two variables, such as blood monocytes(%), and blood eosinophils(%), so $p = 2$, and ten patients, $N = 10$. In the following matrix, each row corresponds to one patient, and the first column presents the recorded values for

blood monocytes(%), while the second column contains the blood eosinophils(%) values. Thus, for example, the observed value of blood eosinophils(%) variable for the fifth patient is $x_{52} = 5.1$.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \\ \mathbf{x}_7 \\ \mathbf{x}_8 \\ \mathbf{x}_9 \\ \mathbf{x}_{10} \end{bmatrix} = \begin{bmatrix} 4.1 & 4 \\ 6.9 & 0.6 \\ 14.5 & 4.3 \\ 5.2 & 6.6 \\ 10.4 & 5.1 \\ 13 & 5.7 \\ 5.6 & 4.6 \\ 7.1 & 0.4 \\ 9.5 & 6.6 \\ 7.7 & 0.7 \end{bmatrix} \quad p = 2; N = 10.$$

On a two-dimensional graph, it will be easy to identify the number of appropriate clusters and also how to classify this data. Figure 2.1 depicts the data values for the 10 patients. Based on this figure, the data can be categorized into three groups where each group is composed of patients with the closest distance to each other.

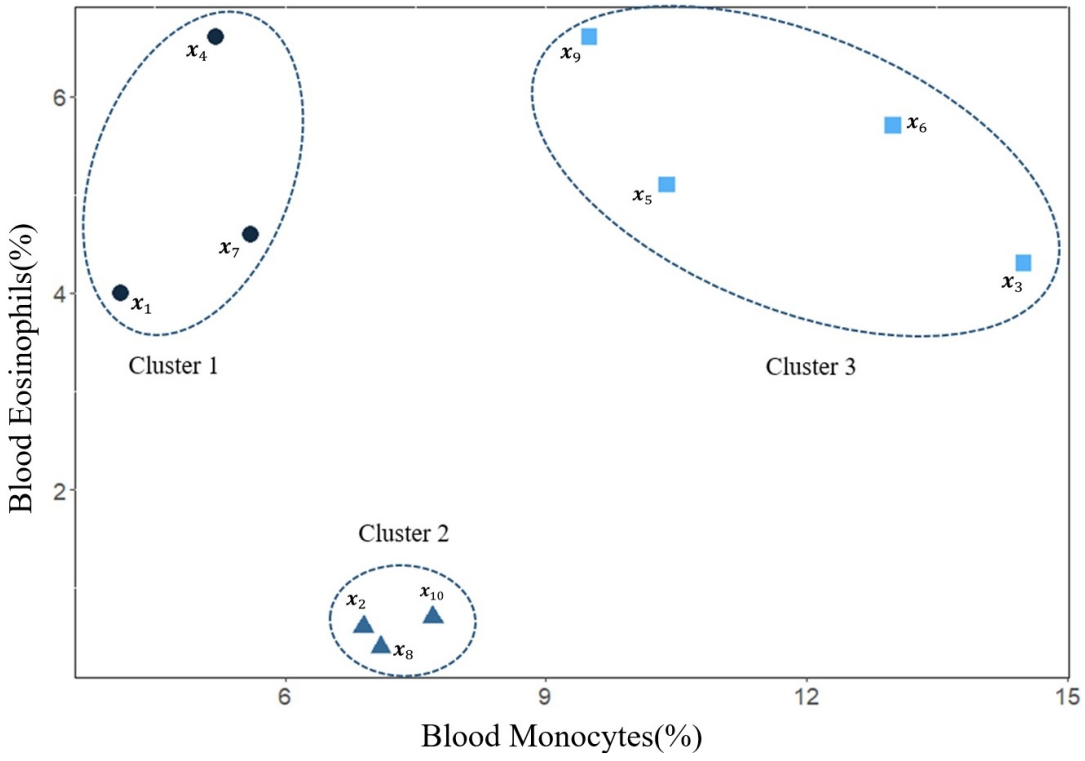

Figure 2.1: An example of clustering

Thus, in previous example, object $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{10}]^T$ is partitioned into $k = 3$ clusters and the result of the clustering can be defined as

$$C_1 = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_4 \\ \mathbf{x}_7 \end{bmatrix}, C_2 = \begin{bmatrix} \mathbf{x}_2 \\ \mathbf{x}_8 \\ \mathbf{x}_{10} \end{bmatrix}, C_3 = \begin{bmatrix} \mathbf{x}_3 \\ \mathbf{x}_5 \\ \mathbf{x}_6 \\ \mathbf{x}_9 \end{bmatrix}.$$

This clustering result follows the three natural conditions listed above. Finally, the corresponding clustering is equal to

$$c = (1,2,3,1,3,3,1,2,3,2),$$

where for example c_5 is a clustering label for the fifth patient, with $c_5 = 3$ meaning that the fifth patient is classified in the third cluster. In order to enable the reader to visually realize the clustering and to get the clustering results, the concept was explained in terms of the easiest $2D$-dataset. However, effective visualization would be impossible for large multidimensional datasets (e.g., more than three variables). Moreover, the classification of clusters using available visualization tools is a difficult task and is not identified straightforward even for lower-dimensional spaces.

In addition, there are other types of clustering methods, which permit objects to belong to more than one cluster simultaneously. Algorithms of this type are commonly referred to as soft algorithms which are not studied in this study.

The process of clustering involves grouping objects according to their similarity or dissimilarity. Therefore, the distance measure plays a crucial role in clustering methods. The majority of clustering methods are directly or indirectly based on distance measures. Accordingly, in order to get the clustering for $\mathbf{X}$, proper distance functions are defined in the next section.

2.1.2 Distance measurements

The most well-known methods of clustering are based on the similarity between two objects $(\mathbf{x}_1, \mathbf{x}_2)$. The distance measure is a function that takes two input objects and returns a real positive number which indicates similarity or dissimilarity between the two objects. The distance measures are the initial step and have a substantial impact on clustering results. Choosing the distance measure depends on the

type of variable and how would be the shape of clustering and, therefore, have an effect on the result of the clustering. In the following, several common distance measures are discussed.

Euclidean distance

The most popular measure of dissimilarity in quantitative data is the Euclidean distance. The Euclidean distance between two p-dimensional data objects, $\mathbf{x}_i = [x_{i1}, ..., x_{ip}]$ and $\mathbf{x}_l = [x_{l1}, ..., x_{lp}]$ is defined as the square root of the sum of squared differences that apply to quantitative variables and is denoted as $d_{Euc}(\mathbf{x}_i, \mathbf{x}_l)$. Based on notations introduced previously, this distance is given by the following equation:

$$d_{Euc}(\mathbf{x}_i, \mathbf{x}_l) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - x_{lj}\right)^2},$$

the Euclidean value gets higher when the two objects are farther apart and then shows the data objects are dissimilar. The lower Euclidean value indicates how similar and close two objects are. Two objects are completely similar and identical when the corresponding Euclidean distance is equal to zero.

Manhattan distance

Manhattan distance, also called rectilinear distance or L_1 distance, measures dissimilarity in quantitative variables and is defined mathematically as:

$$d_{Man}(\mathbf{x}_i, \mathbf{x}_l) = \sum_{j=1}^{p} \left|x_{ij} - x_{lj}\right|.$$

When Manhattan distance is used, the clusters tend to form rectangle shapes. The advantage of Manhattan distance is that it takes less time to compute (Jain and Dubes, 1988). Manhattan distance values can also be interpreted like Euclidean distance values.

Mahalanobis Distance

According to Mahalanobis, different patterns can be identified by considering the correlations between quantitative variables. The squared Mahalanobis is defined as follows:

$$d_{Mah}(\mathbf{x}_i, \mathbf{x}_l) = \sqrt{(\mathbf{x}_i - \mathbf{x}_l)\, \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_l)^T},$$

where Σ is covariance matrix calculated in $\Sigma = \frac{1}{m} \sum_{i=1}^{n} \left(\mathbf{x}_i - \mu\right)\left(\mathbf{x}_i - \mu\right)^T$ with $\mu = \frac{1}{m} \sum_{i=1}^{n} \mathbf{x}_i$ being the vector of average values. Σ^{-1} is inverse of the covariance matrix and presents

how tightly clustered the objects are around the mean. Unlike Euclidean distance, it considers the correlations between the datasets. In general, a lower Mahalanobis value indicates more similarity. Clusters based on the Mahalanobis distance tend to be ellipsoidal and are useful in the identification of the outliers (Mahalanbois, Mahalanbois).

Cosine similarity

Cosine similarity is a metric used to determine the similarity between two non-zero objects. It is calculated as

$$d_{cos}(\mathbf{x}_i, \mathbf{x}_l) = 1 - \frac{\mathbf{x}_i \mathbf{x}_l}{\|\mathbf{x}_i\| \, \|\mathbf{x}_l\|} = 1 - \frac{\sum_{j=1}^{p} x_{ij} * x_{lj}}{\left\| \sum_{j=1}^{p} x_{ij}^2 \right\| \left\| \sum_{j=1}^{p} x_{lj}^2 \right\|}.$$

Cosine similarity ranges between 0, and 1. Objects with cosine values of 1 have the most similarity, while those with zero have the least similarity (Hamerly, 2003).

Gower's Distance

The values of objects can be expressed in different scales (numerical, nominal, or ordinal). However most practical and theoretical results have been presented under the assumption that object components are numerical. A general and useful method for measuring the distance between two objects that may include logical, categorical, numerical, or text data is Gower's Distance. The distance is always a number between 0 and 1, where 1 represents maximum similarity and 0 is being exactly dissimilar. The Gower's distance can be defined as

$$d_{Gow}(\mathbf{x}_i, \mathbf{x}_l) = \frac{\sum_{j=1}^{p} \delta_{ilj} * s_{ilj}}{\sum_{j=1}^{p} \delta_{ilj}},$$

where δ_{ilj} provides the ability to make comparisons, if x_i and x_l can be compared for the j^{th} variable $\delta_{ilj} = 1$, if x_i and x_l cannot be compared, for example because of missing values, δ_{ilj} is set to be zero. The value for s_{ilj} depends on the type of variables, for continuous variables, $s_{ilj} = 1 - \frac{|x_i - x_l|}{R_k}$ where R_k is the range of j^{th} variable. For categorical variable, $s_{ilj} = I\{x_i = x_l\}$ which is equal to 1 if two objects x_i and x_l have the same value in j^{th} variable and $s_{ilj} = 0$ if they are different (Gower, 1971).

Distances between objects can be computed using several R functions, including *dist()* in stats package, *get–dist()* which is implemented in *factoextra()* package, and *daisy()* in cluster package.

2.2 Clustering Techniques

Various clustering methodologies have been presented in the literature, each based on a different induction principle. It is due to the fact that the clustering notion is not clearly defined and leads to many introduced clustering methods (Estivill-Castro, 2002). However, throughout this thesis, we have considered three common clustering methods, which are detailed in following sub-sections. A summary of the advantages and disadvantages of these techniques is summarized in Table 2.1.

Table 2.1: Summary of Clustering Techniques

Techniques	Advantage	Disadvantage
Partition clustering	Straightforward to implement; Guarantees convergence	Don't handle missing values; Predetermine number of clusters; Sensitive to outliers; Dependent on initial values
Hierarchical clustering	Handling missing values by using the distance matrix; No need to predetermine the number of clusters.	Don't handle outliers well; Mistakes made during clustering cannot be reversed
Model-based clustering	Take into account the distribution of datasets; Take into account variance in clustering	Don't handle missing values; Time-consuming; Hard to estimate the parameters

2.2.1 Partition clustering

By definition, the partitioning methods attempt to separate and homogenize objects to perform clustering by optimizing a predefined objective function. These methods relocate objects between clusters until an optimal local partition is achieved by minimizing or maximizing a numerical objective function. Therefore, when objects are grouped together, they are as close as they can be to each other (intra-cluster compactness), and they are well separated from other objects in other clusters (inter-cluster separation). Throughout this section, common methods and algorithms in the partition clustering field are discussed.

K-Means method

K-means method is one of the oldest and simplest cluster analysis (MacQueen, 1967). In this method, the number of clusters, K, and their centroids are defined in advance. Cluster centroid refers to the object that symbolizes the cluster's center. Each object is assigned to one cluster based on the nearest point to the cluster centroid and no

clusters are left empty. The number of clusters, K, is a priori unknown and has to be determined either by the user or by using several different methods for selecting the best number of clusters.

K-means method is an iterative process which begins by randomly assigning K points as cluster centroids. In the next step, objects are assigned to the nearest centroid. The nearest is defined, according to the type of variable, by one of the metric distances described in Section 2.1.2. Once the clusters have been created, the centroid of the updated clustering is computed by the mean of the updated objects in the cluster, $\mu = \frac{1}{n_k} \sum_{i \in C_k} x_i, k = 1, ...K$. In the final step, the previous two steps are repeated until a local minimum of the following within-cluster sum of squares is reached.

$$L_{K-Mean} = \min_{\{C_1,...,C_k\}} \min_{\{\mu_1,...,\mu_k\}} \sum_{k=1}^{K} \sum_{i \in C_k} d(x_i, \mu_K).$$

A description of this iterative algorithm is presented in Table 2.2.

Table 2.2: Algorithm of K-means method

Input: Matrix **X** representing a set of objects, number of clusters k.
Step 1. Determine K points into the data space for objects. These points define the starting cluster of centroids,
Step 2. The distance between each object and centroids is calculated based on their type and objects are assigned to the closest centroid,
Step 3. Recalculate the locations of the K centroids. The new location is calculated by taking the mean of all objects that are assigned to that centroid's cluster,
Step 4. Steps 2 and 3 should be repeated until the centroids are no longer changing locations and the distance between objects and new centroids is as small as possible.
Output: Partition of clustering labels $C = \{C_1, \ldots, C_k\}$

The *stats* R package can be used to compute and visualize partitioning clustering by K-means. The *kmeans()* function is the standard R function to perform K-means clustering. The function can only be applied to the raw datasets. In this function, possible values for the number of clusters should be defined. There were iterations on assigning objects to clusters until no more change occurred. However, the maximum number of iterations can be selected in the function, and the default is 10.

Partition around medoids (PAM) method

The PAM method performs similarly to K-mean method. In K-means, the initial values for centroids are determined randomly, so, results may differ according to the initial values. In the PAM method, objects are assigned to clusters centroids and dissimilarities between a point in the cluster and the point labeled as the cluster's center are minimized. So, the aim of PAM method is to minimize the following function

$$L_{PAM} = \min_{\{C_1,\ldots,C_k\}} \min_{\{\mu_1,\ldots,\mu_k\}} \sum_{k=1}^{K} \sum_{i \in C_k} d(x_i, x_{(k)}),$$

where $x_{(k)}$ represents an object considered as a medoid for k^{th} cluster, C_k. A medoid is a centroid chosen from objects in a dataset. In the first step, the object, $x_{(k)}$, with presents the smallest sum of similarities to other objects is selected as centroids of each cluster. The process is iterated until k objects are selected. All of the selected $x_{(k)}$ are considered as the initial k medoids for each of the k clusters. After this step, according to their proximity to a medoid, remaining objects are grouped into clusters. Next, the algorithm attempts to find new medoids in each cluster based on objects from that cluster whose distance (L_{PAM}) makes a minimum of the sum of distances. The process is repeated until the k medoids are no longer changed (Kaufman and Rousseeuw, 1990). A description of the algorithm is summarized in Table 2.3.

Table 2.3: Algorithm of PAM method

Input: Matrix **X** representing a set of objects or dissimilarity matrix, number of clusters k.
Step 1. Determine randomly K objects for an initial set of medoids, **Step 2.** The distance between each object and the medoids are calculated based on their type and objects are assigned to the closest medoids, **Step 3.** Improve the quality of clustering by exchanging selected and unselected objects, **Step 4.** Steps 2 and 3 should be repeated until the distance between objects and medoids is as small as possible
Output: Partition of clustering labels $C = \{C_1, \ldots, C_k\}$.

Since the medoid vector was chosen from the objects, this method is, unlike K-means, capable of analyzing both dissimilarity matrices and raw data. Therefore, besides numeric datasets, the function can be applied to a dissimilarity matrix derived using one of the distance methods. The *pam()* function in *stats* R package is used to compute and visualize partitioning clustering by PAM method.

Define number of clusters in partition clustering methods

When partition clustering methods are used, the number of clusters should be determined before the clustering process begins. Several methods exist for determining the number of clusters in partition clustering methods. The two most well-known silhouette and elbow methods are illustrated here below.

Silhouette Method

The Silhouette method was proposed by Kaufman and Rousseeuw (1990) to estimate the appropriate number of clusters to consider. The silhouette value indicates how similar an object is to its own cluster in comparison to other clusters. The silhouette value ranges between -1 and +1. In general, a higher silhouette value means a better match between the cluster and the object. The silhouette value can be determined as follows using any distance metric. Let's defined

$$a(i) = \frac{1}{n_i - 1} \sum_{i \neq j, j \in C_i} d(x_i, x_{(j)}),$$

the average of dissimilarity between object x_i and all other objects in the cluster C_i in which x_i is part of and,

$$b(i) = \min_{i \neq j} \frac{1}{n_j} \sum_{j \in C_j} d(x_i, x_{(j)}),$$

the minimum average distance of object x_i from all the objects in another clusters, C_j, except other member of cluster x_i. The silhouette value for an object is defined as

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

It should be noted that $a(i)$ is not defined for a cluster of a single object, and in order to calculate $b(i)$, at least two clusters are required. By measuring the silhouette values for all the objects, if the mean silhouette value is at the highest value, the clustering structure is appropriate. In contrast, if the mean silhouette value became lower or negative, then the cluster structure is not optimal, meaning that either too many clusters or too few clusters are considered.

In order to determine the optimal number of clusters, the mean of silhouette value is plotted against the potential number of clusters (usually 10 clusters). The y-axis of the plot represents the average of Silhouette value, and the x-axis indicates the potential number of clusters. The higher silhouette value indicates a proper number

for clustering. As an example, if we consider just three continuous variables, blood eosinophils(%), monocytes(%), and neutrophils(%) in our dataset on 41 patients. The corresponding plot for silhouette value is presented in Figure 2.2. The plot shows that it is more appropriate to classify patients into two clusters.

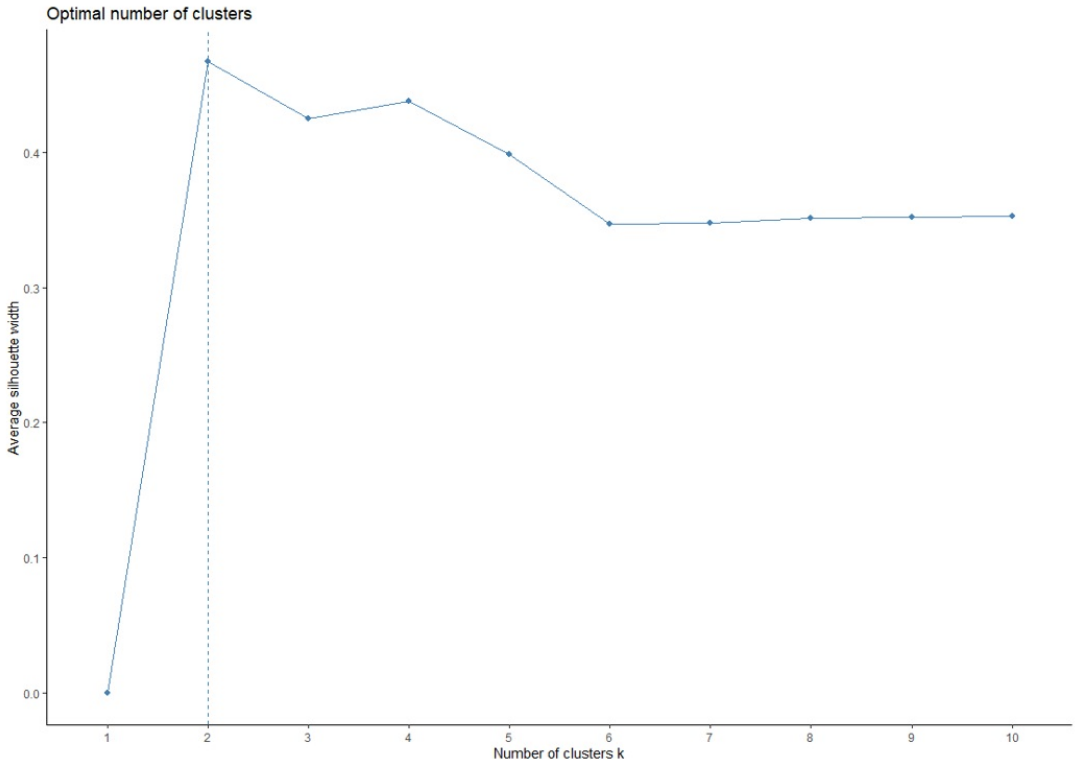

Figure 2.2: Optimal number of clusters using the Silhouette value

Elbow Method

In the elbow method, the sum of squared errors within each cluster is defined

$$SSW = \sum_{k=1}^{K} \sum_{i \in C_k} \left(x_{ik} - \mu_k\right)^2,$$

which calculates the sum of distances between the objects and the corresponding centroids for each cluster, μ_k, and plotted it against the number of clusters. For the first few clusters, there will be a lot of variance and information to explain, but after a certain number of potential clusters, the information will be decreased intangible, which will give the graph an elbow. The plot may seem ambiguous sometimes, and identifying the elbow, we are looking for is quite blind and up to individual discretion. However, this point can be selected as the proper number of clusters. Considering

the same example as above, the plot for sum of squared errors measured within each cluster versus potential number of clusters (usually 10 clusters) is shown in Figure 2.3. The plot suggested that patients could be classified into two clusters.

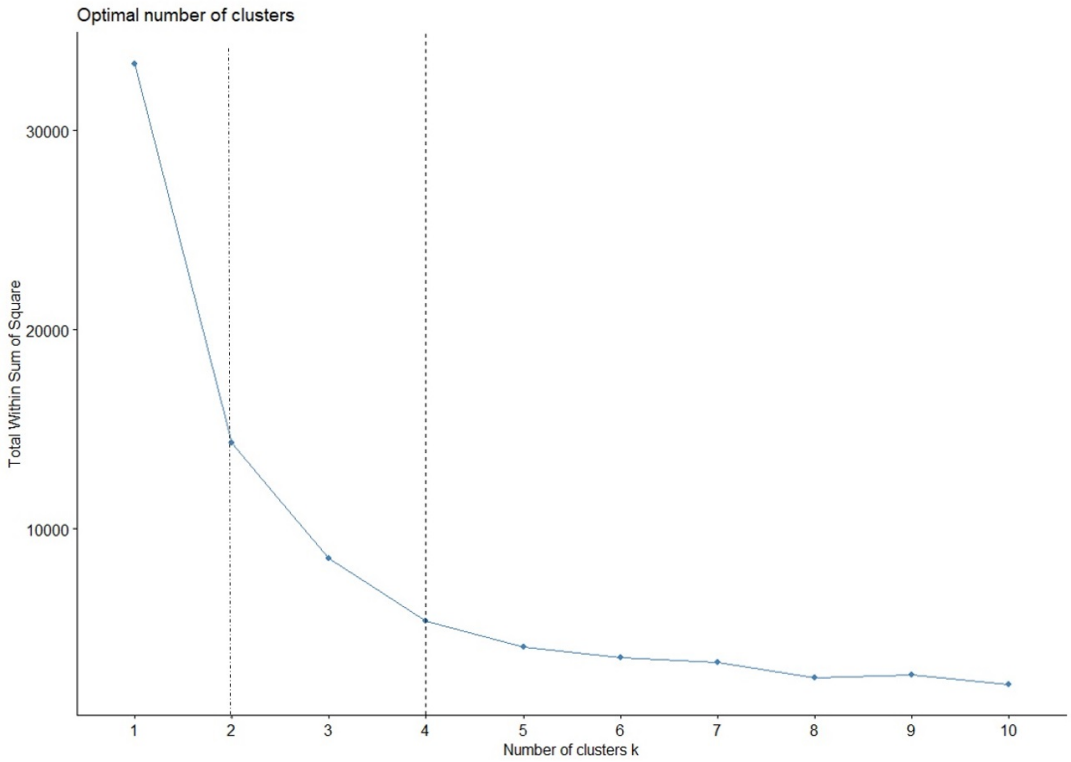

Figure 2.3: Optimal number of clusters using the Elbow method

2.2.2 Hierarchical Clustering

Hierarchical clustering is a well-famous clustering method. The algorithms of hierarchical clustering are quite popular in the literature for their attractive visual outputs. The algorithm consists of two main types, agglomerative and divisive. The agglomerative clustering method creates a single cluster from each object, then a combination of close clusters merges iteratively to create a new cluster, and so on until a single cluster is formed. In contrast, the divisive method involves grouping all the objects into one large cluster and dividing them iteratively into groups that are geographically far apart, until each object was alone in their own cluster. A description of the algorithm for agglomerative hierarchical clustering is presented in Table 2.4.

For example, the first five observations of FEV1 predicted(% predicted) were considered in Figure 2.4 and their corresponding values are shown in the square. A num-

Table 2.4: Algorithm of agglomerative hierarchical clustering method

Input: Matrix **X** representing a set of objects or dissimilarity matrix
Step 1. Objects are clustered into single elements and distances are determined for each pair of clusters.
Step 2. Find the pairs C_i and C_j of clusters that are closest to each other, in purpose of generating dendrogram, this corresponds to inserting a new node and connecting it with the other nodes.
Step 3. Recalculate the distance between C_k and the remaining clusters
Step 4. Steps 2 to 3 should be repeated until only one, single cluster is created
Output: Dendrogram and partition of clustering labels $C = \{C_1, \ldots, C_k\}$

ber from 1 to 5 was assigned to each object, respectively. In agglomerative clustering, each participant is initially viewed as a single cluster in Step 0. A new cluster is created at each step of the algorithm by combining two clusters that are most similar and here, have a lower Euclidean distance from one another. In step 1, we determined that the values 2520 and 2570 had the closest distance and were combined. This process is repeated until all objects belong to a single big cluster. The divisive method works in a right to left manner. It begins with all objects being included in a single cluster. In each iteration, the most heterogeneous cluster is divided into two. The process is repeated until all objects are in their own cluster. The graphic representation of these two methods is called dendrograms (Everitt, 2011).

For these two types of hierarchical methods, after step 0, we need to calculate the similarity or homogeneity between two clusters. Several methods are presented for calculating the homogenous between two clusters, such as the single, complete, average, centroid methods and Ward criterion. Based on one of the following criteria, two clusters are merged in agglomerative method or dispersed in divisive method.

The Single Linkage criterion

In the single linkage method, the dissimilarity between two clusters, C_i,C_j, is defined as the nearest distance between a pair of objects, one object from C_i and one from cluster C_j. According to this method, the smallest distance between two objects of clusters is defined as the degree of dissimilarity between the two clusters.

$$d_{aver-Single}\left(C_i, C_j\right) = \min_{i' \in C_i, j' \in C_j} d(\mathbf{x}_{i'}, \mathbf{x}_{j'}).$$

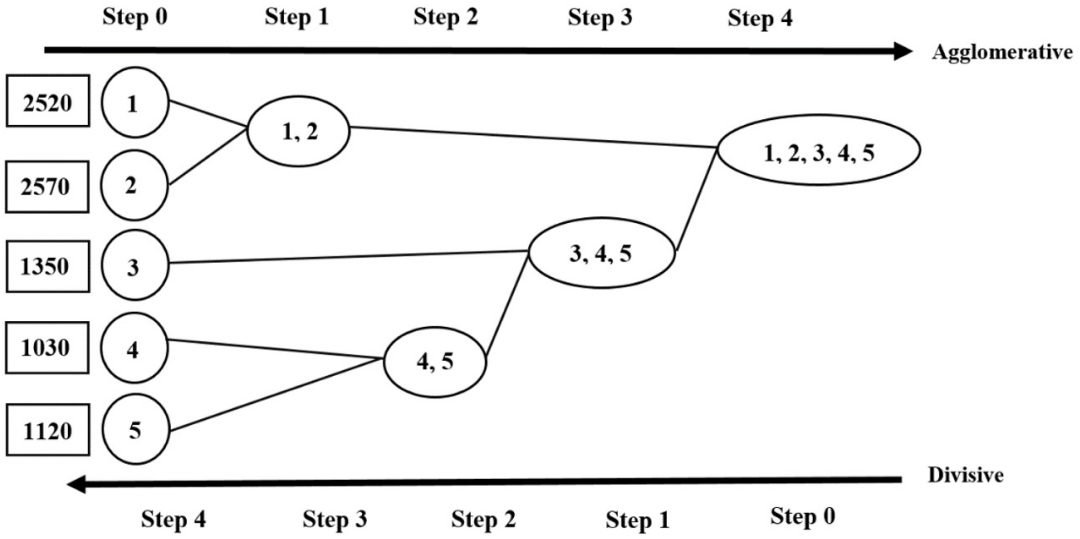

Figure 2.4: Dendrograms of agglomerative and divisive hierarchical clustering

By using this method, the derived clusters are often separated according to their dissimilarity. $d(\mathbf{x}_{i'}, \mathbf{x}_{j'})$ is one of the distance measures that were introduced in Section 2.1.2. The exhibit of this criterion is shown in Figure 2.5(a).

The Complete Linkage criterion

Complete Linkage is typically referred to as the furthest neighbour rule, which results from the largest distance between an object in one cluster, C_i, and an object in another cluster, C_j. In this way, a cluster's dissimilarity is defined as

$$d_{aver-Complete}(C_i, C_j) = \max_{i' \in C_i, j' \in C_j} d(\mathbf{x}_{i'}, \mathbf{x}_{j'}).$$

This method prevents clusters from merging together if there are objects within a cluster that are far apart. These results have compact clusters with similar diameters, which is a benefit of this method. The clusters results, however, may not be well separated (Figure 2.5(b)).

The Average Linkage criterion

In this method, the distance between two clusters is measured by the average of distances between all pairs of objects from two clusters, C_i and C_j. One object in the pair of the distance is from C_i and another object is in C_j. The average linkage is calculated as

$$d_{aver-Linkage}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{i' \in C_i, j' \in C_j} d(\mathbf{x}_{i'}, \mathbf{x}_{j'}),$$

where n_i and n_j are the numbers of objects in the two clusters. The average linkage method is considered to be a compromise between the complete linkage method and the single linkage method. Dissimilarities between clusters are not determined by the behavior of two objects, but rather by the collective behavior of the clusters(Figure 2.5(c)).

The Centroid Linkage Criterion

To calculate the distance between clusters, it is also possible to take into account the difference between their centroids:

$$d_{aver-Centroid}(C_i, C_j) = d\left(\frac{1}{n_i} \sum_{i' \in C_i} \mathbf{x}_{i'} - \frac{1}{n_i} \sum_{j' \in C_j} \mathbf{x}_{j'}\right),$$

Centroid linkage calculates the centroid of each group and then determines the distance between them. Obviously, this criterion only makes sense if an average of objects is reasonable (Figure 2.5(d)).

Ward Linkage Criterion

The general idea of Ward criterion was to evaluate and optimize the objective function when two clusters are combined. A well-known objective function is the variance method or the sum of squares error within clusters that should be minimized when two clusters are merged. As explained in the elbow method, sum of squares error within clusters is defined as follow

$$SSW = \sum_{k=1}^{K} \sum_{i \in C_k} \left(x_{ik} - \mu_k\right)^2,$$

where μ_k is the mean of object in k^{th} cluster. The Euclidean distance method is used here, however, other distances can be applied. It is not necessary to compare all objects to a mean, centroid, medoids, mode, or other commonly used averages can be used instead.

Cluster final outputs are strongly influenced by the linkage criteria considered. Each linkage criterion has its own properties. Each of these criteria performs well under certain circumstances. For instance, single linkage may be used to handle

complex shapes with outliers and is not concerned about compactness. Complete linkage usually results in clusters that are almost identical in size. Ward's method allows to derive clusters with equal sizes and spherical shapes.

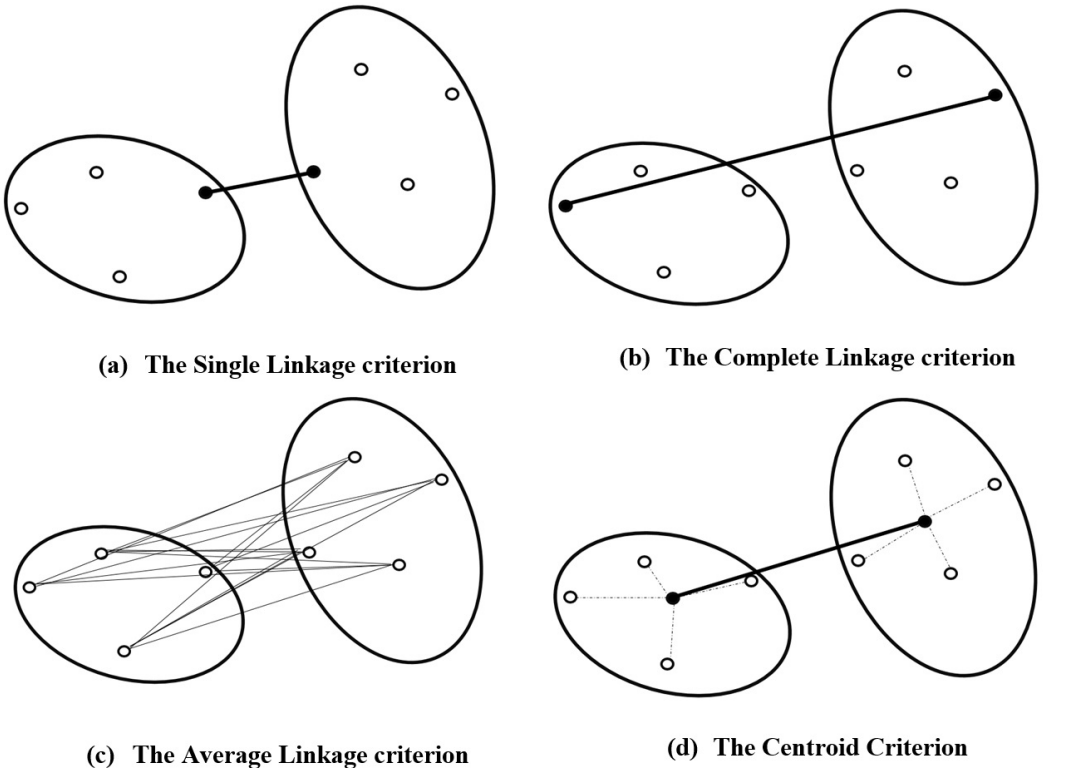

Figure 2.5: Four different types of linkage criteria for hierarchical clustering (Inspired by Figure 17.3 of Introduction to Information Retrieval (Manning et al., 2008))

In contrast to partition clustering, hierarchical clustering creates a hierarchy of clusters and does not require pre-defining the number of clusters. In Figure 2.6, dendrogram for the first 15 observations of FEV1 predicted(% predicted) were considered. Distance (or similarity) based on Ward linkage criterion between the two merged clusters is shown on the Y-axis. It is possible to quickly determine the number of clusters by dendrogram in hierarchical method. On the dendrogram plot, one value can be defined on the Y-axis, a horizontal line should be drawn at that value, this line intersects the horizontal lines in the dendrogram which can be counted as the number of clusters. In Figure 2.6, we can consider two horizontal lines, which define two or four clusters. In the final analysis, the researcher is responsible for determining the number of clusters, however, the highest distance could be considered to determine the final number of clusters. So, in the example, based on the hierarchical cluster, the objects can be classified into two clusters.

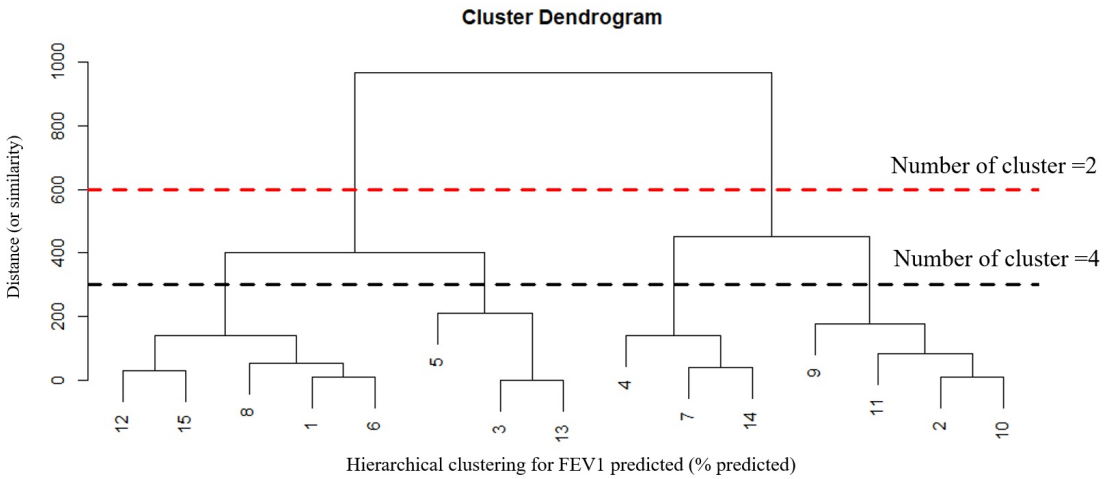

Figure 2.6: Dendrograms in determining cluster size

hclust() in *stats* package and *agnes()* in *cluster* package can be used for hierarchical clustering and for drawing dendrograms in agglomerative method, and *diana()* in *cluster* package can be used for divisive clustering.

2.2.3 Model-based clustering

One of the most effective clustering approaches is model-based clustering which involves the use of a mixture model (McLachlan and Peel, 2000; Zhong and Ghosh, 2003). Model-based clustering enables to make statistical inferences and estimate uncertainty for parameters or clustering assignments. In the model-based clustering approach, objects belonging to the same distribution are grouped together. In this method, data is assumed to follow a certain probability distribution model. The likelihood of the potential model is maximized by estimating mixture model parameters. If the complexity of the model is not constrained, this approach over-fits the dataset. However, model-based approaches are stronger than hierarchical clustering and partition clustering since they provide not only clusters, but also a mixture model, which we can use to better understand the distribution of the data.

Generally, in model-based clustering, clusters are formed based on multivariate Gaussian distributions, then all clusters are combined with corresponding probabilities such that they add up to one. This distribution is called the mixture Gaussian distribution and is known as

$$f(\mathbf{x},\theta) = \sum_{k=1}^{K} \eta_k \phi\left(\mathbf{x}|\mu_k,\Sigma_k\right),$$

where x is an independent sample which issued from K-component mixture distribution with p-dimensional outcome, k indicates a specific cluster and η_k is the mixing proportion of k^{th} cluster with $\eta_k > 0$ and $\sum_{k=1}^{K} \eta_k = 1$. The parameters θ in this model are the mixing proportions of the mixture, η_k, as well as the cluster specifics parameters of Gaussian distribution which are the cluster means μ_k and the cluster covariance matrices Σ_k for $k = 1, \cdots, K$.

$\phi(.|.)$ is the probability density function of the multivariate Gaussian distribution with mean μ_k and variance-covariance matrix Σ_k for each determined clusters,

$$\phi_k\left(\mathbf{x}|\mu_k,\Sigma_k\right) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}-\mu_k\right)\right),$$

corresponding to the k^{th} cluster.

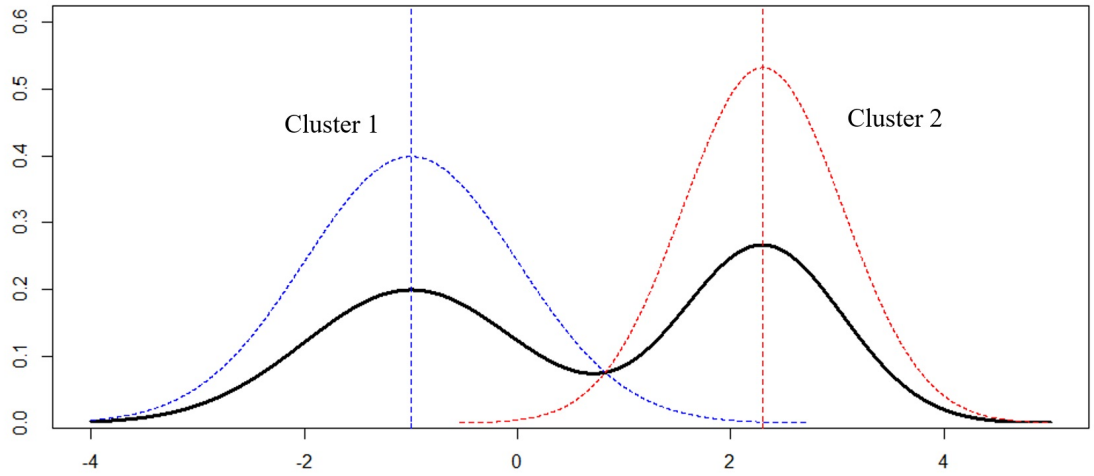

Figure 2.7: Density of a one-dimensional Gaussian mixture distribution with two components

Suppose that the purpose of the study was to classify patients according to just one variable. In the plot for this variable, the black solid line in Figure 2.7, it is obvious that the sample will consist of two clusters, each with a normal distribution. In model-based clustering, we determined the number of clusters, and then estimated the probability of these mixed clusters as well as the parameters of these two normal distributions (Dashed lines in red and blue).

Whenever the number of variables increases and there are multiple types of variables in the dataset, advanced methods are required to estimate the parameters and determine the number of clusters. The variance-covariance matrix, Σ_k, will be parameterized in order to make the model more adaptable to these types of datasets:

$$\Sigma_k = \lambda_k D_k A_k D_k^T.$$

where $\lambda_k = |\Sigma_k|^{1/d}$ measures the volume of the k^{th} cluster. D_k is the matrix of eigenvectors of Σ_k which determines its orientation, and A_k determines the shape of model. A_k is a diagonal matrix containing the normalized eigenvalues of Σ_k in decreasing order along the diagonal, and $|A_k| = 1$.

If we could impose more restrictions on parameters in the model, the diagonal family and spherical family may be suggested. In addition, eight general models were proposed based on this variance matrix decomposition, which is shown in Table 2.5.

Table 2.5: Different types, names, and structures for variance-covariance matrix

	Model	Volume	Shape	Orientation	Σ_k
Spherical	EII	Equal	Spherical		λI
	VII	Variable	Spherical		$\lambda_k I$
Diagonal	EEI	Equal	Equal	Axis-Aligned	$\lambda\Delta$
	VEI	Variable	Equal	Axis-Aligned	$\lambda_k\Delta$
	EVI	Equal	Variable	Axis-Aligned	$\lambda\Delta_k$
	VVI	Variable	Variable	Axis-Aligned	$\lambda_k\Delta_k$
General	EEE	Equal	Equal	Equal	$\lambda D\Delta D^T$
	VEE	Variable	Equal	Equal	$\lambda_k D\Delta D^T$
	EVE	Equal	Variable	Equal	$\lambda D\Delta_k D^T$
	EEV	Equal	Equal	Variable	$\lambda D_k\Delta D_k^T$
	VVE	Variable	Variable	Equal	$\lambda_k D\Delta_k D^T$
	VEV	Variable	Equal	Variable	$\lambda_k D_k\Delta D_k^T$
	EVV	Equal	Variable	Variable	$\lambda D_k\Delta_k D_k^T$
	VVV	Variable	Variable	Variable	$\lambda_k D_k A_k D_k^T$

Once the model has been defined, the next step is to estimate the mixture model parameters by maximizing the likelihood, and determining the data partition from the estimated parameters. Expectation-Maximization (EM) algorithm can accomplish this step. In model-based clustering, the clustering results are considered as missing values, so EM algorithm is an efficient method for computing the Maximum Likelihood (ML) estimate by maximizing the expectation of complete log-likelihood.

Therefore, the cluster of each object is unknown and is indicated by $\mathbf{z}$, $\mathbf{z} = (z_1, z_2, \cdots, z_n)$, where $z_i = k$ when i^{th} object, x_i, belongs to the k^{th} clusters. The complete log-likelihood can be presented as follows:

$$CL(\mathbf{z}, \theta|\mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log\left(\eta_k \phi(\mathbf{z}, \theta|\mathbf{x})\right).$$

The EM algorithm begins with a set of initial parameters $\theta^{(0)}$, then iterates between the expected step (E step) and the maximization step (M step).

E step calculates the expectation value of the complete log-likelihood function with respect to the conditional distribution of z given x under the current estimate of the parameters θ;

$$Q\left(\theta|\theta^q\right) = E\left[\log CL(\mathbf{z}, \theta|\mathbf{x})\right],$$

then calculating the posterior probabilities $\pi_{ik}^{(q)}$ of x_i belonging to the k^{th} clusters;

$$\pi_{ik}^{(q)} = \frac{\pi_k^{(q)} \phi_k\left(x; \theta_k^{(q)}\right)}{\sum_l \pi_l^{(q)} \phi_l\left(x; \theta_l^{(q)}\right)}.$$

In M step, the parameter $\theta^{(q+1)}$ that maximize the expectation in the previous E step will be found by

$$\theta^{(q+1)} = \arg\max_{\theta} Q\left(\theta|\theta^q\right).$$

Depending on the quantity of data and the dimensionality of variables, convergence might require many iterations and a long computation time.

In fitting the model and estimating the parameters, the number of clusters is a crucial factor in cluster analysis. It is also important to consider the type of fitted model and the kind of variance-covariance decomposition in this method. Although, we might deduce the potential model and number of clusters from a simple distribution with one or two-dimensional variables, however, if the variables in the dataset are multidimensional and of various types, or if they are well-mixed, we need to introduce the necessary criteria. Due to the fact that this method must identify the number of clusters and type of the model simultaneously, the model criteria will be quite sensitive.

Experimentally, the model that best represents the data distribution and could

adapt better to the characteristics of the data, produce the best results of the mixture models in clustering and increases the maximum likelihood. However, if the number of clusters is so close to the number of objects, the clustering concept is lost. Therefore, maximum likelihood alone is not a good criterion for choosing the number of clusters. It is important to establish a balance between data information and model parameters in the criteria. The information criteria, Akaike information criterion (AIC), and Bayesian Information Criterion (BIC), are commonly applied in fitting models that are based on maximum likelihood and penalized by the number of model parameters.

As a first criterion for model selection, AIC is generally regarded as the appropriate option:

$$AIC = -2L_{\max}\left(\hat{\theta}\right) + 2\nu\left(\theta\right),$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ, $L_{\max}\left(\hat{\theta}\right)$ is the maximum of log-likelihood for the estimated model:

$$L_{\max}\left(\hat{\theta}\right) = \max \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{K} \eta_k \phi\left(\mathbf{x}|\mu_k, \Sigma_k\right)\right),$$

and $\nu\left(\theta\right)$ is the number of the free parameters in the model. Another penalized criterion that is highly related to AIC is the BIC.

$$BIC = -2L_{\max}\left(\hat{\theta}\right) + \nu\left(\theta\right)\ln(n),$$

where n is a number of sample size. Comparing competing models can be carried out using both criteria, however the BIC criterion is preferred when the models have different parameters or different numbers of clusters. Among competitive models, the model with the highest AIC or BIC illustrates the best model for the number of clusters and also clustering results. The general rule in clustering in contrast to modeling is that a large BIC or AIC value indicates strong evidence for the model and defined number of clusters.

For example, if we consider three continuous variables, blood eosinophils(%), monocytes(%), and neutrophils(%) in our dataset. All missing values and outliers were removed, then variables were normalized. On 41 selected patients, model-based clustering was performed and a search of all 14 types of models and $1-9$ clusters was conducted. Figure 2.8 illustrates the BIC scores and shows that the best clustering

result comes from VVV type of model with three clusters.

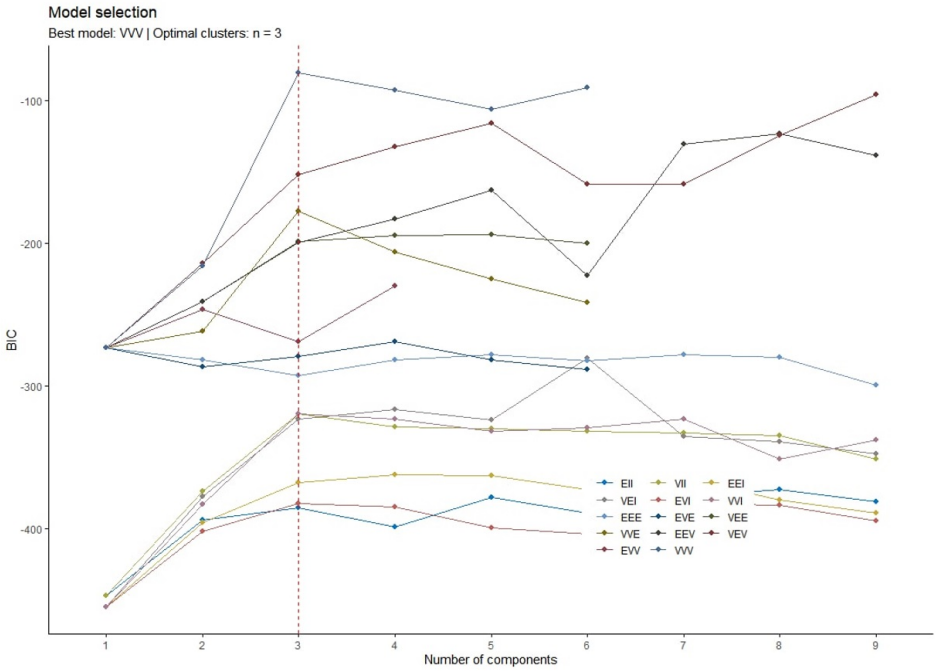

Figure 2.8: Identifying the optimal model-based clustering and number of clusters based on BIC

In some models, results are not generated for all cluster sizes (e.g., the VVV model produces results for clusters $1-6$). Models that are applied to these settings do not converge to optimal results. This becomes increasingly problematic as datasets become larger. It is often beneficial to perform dimension reduction before performing a model-based clustering.

In this section, we explained model-based clustering for normal distributions. This method was able to extend to all types of variables and their distributions and could also be applied to mixed datasets. Fitting the model, estimating the parameters, and comparing and choosing the models are all handled using *mclust* package and function in R.

2.3 Clustering Validation

Cluster analysis is a powerful tool for finding structure in a dataset. However, it is also an unsupervised method which means the number of clusters, K, is unknown and there is no prior knowledge of it. In addition, there is also no information of true

cluster or obvious grouping in the dataset. Therefore, one of the major challenges in this unsupervised method is evaluating the accuracy, quality, and goodness of the clustering framework for classification. The main objective of cluster validation is to determine which partitioning of the data best matches the underlying data. The validation issue of cluster analysis is more challenging than evaluating the quality of regression, predictions, or other supervised models. Indeed, true cluster and grouping information is typically not available. For validation of the clustering, three statistical methods, internal, external, and cluster stability, can be considered.

Validating clustering results aims to determine which results are most meaningful and to compare multiple clustering methods. Therefore, the object of cluster validating and estimating the number of clusters comes the closely same task. Cluster validation can be done by using several of the clustering quality indices that are reported for estimation of the number of clusters. Similarly, the number of clusters can be selected by optimizing one or several indexes in cluster validation.

2.3.1 Internal Validation

Internal indices evaluate the quality of a cluster based on the clustering and the underlying dataset. The interval validation tools are designed to evaluate the clustering concept that groups similar objects within the same cluster and clusters dissimilar objects together. Therefore, the two concepts of intra-cluster similarity and inter-cluster similarity are applied here. In terms of similarity between objects, intra-cluster similarity (compactness, connectedness, and homogeneity) measures how similar the objects within a cluster are, and the distance between clusters is measured by inter-cluster similarity or separation. An effective clustering of a dataset is the clustering that provides maximum separation in the dataset. There are different ways to determine separation. Methods include calculating the distance between the closest objects, the most distant objects, and the centers of two clusters. The majority of these indices are also introduced to determine the number of clusters which has the best fit the data.

Sum of squares within clusters (SSW)

The SSW metric measures cluster compactness and is based on the centroid of clusters. It is particularly suitable in cases where hyper-spherical clusters are desired. Due to its dependency on cluster centroids, c_k, this index is only applicable to numerical data. As the number of clusters increases, the value of SSW decreases.

$$SSW_k = \sum_{i=1}^{N} (x_{ik} - c_{ki})^2.$$

In general, a classification that has a lower calculated SSW is preferable. However, if there are many clusters, this value will automatically decrease, which is undesirable.

Sum of square between clusters (SSB)

Sum of squares between clusters (SSB) is a measure of separation between clusters based on the variance between clusters. In order to separate clusters, the distances from the centroids to the mean vector of all objects are calculated. Based on the calculation, a larger cluster has a greater effect on the index. The SSB value usually increases as the number of clusters increases.

$$SSB_k = \sum_{k=1}^{K} n_k \| c_k - \bar{x} \|^2.$$

Calinski-Harabasz index (CH)

To provide the best separation and compactness simultaneously, the Calinski-Harabasz (CH) index takes into account the ratio of separation and compactness (Calinski and Harabasz, 1974). When the index value is the maximum, the clustering has a high separation and is compact. If a dataset contains more clusters, the SSB will be higher and the SSW lower. However, the decrease in SSW is greater than in SSB. Thus, by imposing a penalty factor of $(K-1)$, it prevents the conclusion of a higher number of clusters than the one correct. $N-K$ is a term used in cases where the number of clusters is comparable to the number of objects. A good value for K is much lower than N, so the term tends to N.

$$CH = \frac{\frac{SSB}{K-1}}{\frac{SSW}{N-K}}.$$

Silhouette Coefficient (SC)

This index is the same Silhouette index that defines the number of clusters. Silhouettes coefficients (SC) assess how well each object fits in its cluster and is separated from objects in other clusters.

$$SC = \frac{1}{N} \sum_{i-1}^{N} \frac{b(x_i) - a(x_{i)}}{\max(a(x_i), b(x_i))}.$$

According to this formula, the average dissimilarity between one object, x_i, with all objects in the same cluster is calculated as $a(x_i)$, which illustrates how well x_i is placed within correct group. Also, $b(x_i)$ is calculated as the lowest average dissimilarity of x_i to other clusters. When calculating the index, the similarity between two objects is the only factor considered. In other words, SC can be used to evaluate the cluster in a wide range of data and to handle any clustering structure. The silhouette width indicates the average degree of confidence in the clustering. Silhouette width falls in the interval $[-1,1]$ and values close to 1 present well clustered.

Dunn family of Indices

Dunn Index divides the smallest distance between two objects in the different clusters to the largest intra-cluster distance

$$DI = \frac{\min_{K\, i=1} \min_{K\, j=i+1} d(c_i, c_j)}{\max_{k=1} diam(c_k)}.$$

where $d(c_i, c_j)$ present the dissimilarity between two clusters,

$$d(c_i, c_j) = \min_{x \in c_i, \acute{x} \in c_j} \sum_{i=1}^{N} \left(x_{ic_i} - \acute{x}_{ic_j} \right)^2,$$

and $diam(c_k)$ is determined by maximum dissimilarity between two objects in one cluster,

$$diam(c_k) = \max_{x, \acute{x} \in c_k} \sum_{i=1}^{n_k} (x_i - \acute{x}_i)^2, x_i, x_j \in c_k.$$

Dunn Index is in the interval $[0, \infty]$ and the maximum values are preferable with better cluster separation and compact clusters. Dunn index has a high time complexity, and is sensitive to noise and outliers. Three related indices, Dunn-like indices, have been added to Dunn index as a solution to these limitations (Dunn, 1974).

2.3.2 External Validation

External validation can be used to determine whether a clustering result is valid based on a predefined clustering's result. A true clustering normally is not available. However, the predefined clustering result could be a clustering that was produced using another method, or an external result that researchers verified as a final result of the clustering process. This strategy will provide insight into the performance of clustering by evaluating how well different clustering methods agree about a given dataset.

Rand Index (RI)

The purpose of Rand's Index is to compare a classification scheme with a correct classification. Typically, the measure is the percentage of correctly classified elements compared to the total number of elements. Thus, the Rand Index is defined by

$$RI\left(C_i, C_j\right) = \frac{2(a+b)}{N(N-1)},$$

where a indicates the number of pairs of objects in both partitions C_i and partition C_j that are in the same cluster, b indicates the number of pairs of objects that are in the same cluster in partitions C_i but in different clusters in partition C_j. When RI is 0, two clusters were classified differently and 1 means identical classification (Rand, 1971).

Adjusted Rand Index (ARI)

Inspired by the Rand Index, which incorporates chance agreements, ARI was developed and is a common method of external validation in clustering literature.

$$ARI = \frac{a - \dfrac{(a+c)(c+d)}{(a+b+c+d)}}{\dfrac{(a+c)(a+b)}{2} - \dfrac{(a+c(a+b)}{(a+b+c+d)}},$$

where c indicates the number of pairs of objects that are in the different clusters in partitions C_i but in the same cluster in partition C_j; finally d indicates the number of pairs of objects that are in the different clusters in both partitions C_i and in partition C_j. ARI range from 0 to 1, and a higher value indicates greater similarity (Hubert and Arabie, 1985).

2.3.3 Clustering Stability Validation

Finally, in stability measure, the consistency of a clustering result is evaluated by comparing it to the data obtained after removing each variable one at a time. Following are some important criteria for clustering stability validation.

Average proportion of non-overlap (APN)

APN measure is defined as;

$$APN(C) = \frac{1}{NP}\sum_{i=1}^{N}\sum_{l=1}^{P}\left(1 - \frac{n\left(C^{i,l} \cap C^{i,0}\right)}{n(C^{i,0})}\right),$$

where $C^{i,0}$ is the cluster for i^{th} object, x_i based on all data, $C^{i,l}$, the result of clustering when the l^{th} column is removed from the data. The APN is between $[0,1]$, values close to zero indicate a highly consistent clustering.

Average distance (AD)

AD calculates the average distance between objects placed in the same cluster by clustering on the full data and clustering after removing one column. AD is defined as:

$$AD(C) = \frac{1}{NP}\sum_{i=1}^{N}\sum_{l=1}^{P}\frac{1}{n(C^{i,0})n(C^{i,l})}\left(\sum_{i\in C^{i,0}}\sum_{j\in C^{i,l}} dist\left(x_i, x_j\right)\right).$$

The AD has a value between zero and ∞, and it is preferable to have a smaller value.

Average distance between means (ADM)

Using clustering both for the full dataset and clustering when omitting a single column, this measure calculates the mean distance between centers of objects within the same group, it is defined as

$$ADM(C) = \frac{1}{NP}\sum_{i=1}^{N}\sum_{l=1}^{P} dist\left(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}}\right).$$

The ADM also has a value between zero and ∞, and a smaller value is preferred here as well.

Figure of merit (FOM)

The FOM measures the average intra-cluster variance of the objects in the removed column and the clustering of the objects in the remaining columns. FOM for the l^{th} deleted column is calculated as follows

$$FOM(l, C) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{\in C_k(l)} dist\left(x_{i,l}, \bar{x}_{C_k(l)}\right)},$$

where $\bar{x}_{C_k(l)}$ is the average of the k^{th} cluster when l^{th} column is removed. The final score is calculated as an average of all the omitted columns.

$$FOM(C) = \sum_{l=1}^{P} FOM(l, C).$$

FOM represents the average compactness of the clustering result. This value ranges from 0 to 1. A low value indicates better discrimination in clustering. Additionally, FOM decreases when the number of clusters increases (Yeung et al., 2001).

2.4 Conclusion

Three commonly applied clustering techniques are described in this chapter, as well as, the method for determining the number of clusters. Different criteria were also presented for cluster validation. In the clustering methods presented so far, there has been no consideration of the potential presence of missing data or the impact of a large number of variables in the dataset. In fact, until now, the problem of missing values was handled by either discarding objects with at least one missing observation or considering zero in distance measurements. There is no evidence to support the number of variables that should be considered for cluster analysis. Despite the fact that the introduced methods may be applied to numerous variables, it is highly recommended not to include too many variables. It is evident that both of these approaches can lead to biased results. Missing data can be classified in several ways, and appropriate methods for dealing with missing data, as well as how to take into consideration a large number of variables, will be discussed in the next chapter.

CHAPTER 3

Challenging of cluster analysis

In order to perform efficient cluster analysis, it is important beforehand to carefully explore the datasets. The presence of missing values is one of the most critical aspects that could have a great impact on cluster analysis. In presence of missingness, it is crucial to understand what kinds of missing values occur and what is the most effective method to deal with them. This issue will be covered in Section 3.1. The multiple imputation process, which is one of the most effective methods to handle missing values, will be detailed. As introduced in Chapter 2, cluster analysis can be heavily influenced by the recorded variables and its application may fail if the variables are noisy or collinear. The purpose of Section 3.2 is then to address this issue. To conclude this Chapter, since multiple imputation was considered in this thesis, consensus clustering should be used to combine the clustering results. In Section 3.3, two common consensus methods are discussed, and then, the proposed method of this thesis for consensus clustering, 4M, is presented.

3.1 Missing Value

Unobserved values in a dataset are typically referred to as missing values. As values can be missing for a variety of reasons, there exist different types of missingness. In the huge datasets considered in this thesis, missing value is a common and pervasive problem. A common reason for occurring missing is that objects do not respond to one or more questions either because they refuse to answer, do not understand the

questions, do not know the answers, or accidentally skip the questions in a survey, or case report form. In the experimental part of studies, missing values can occur when a researcher is unable to collect an observation. In some other cases, poor environmental conditions or patient circumstances may render observation impossible, for example, when patients are not able to pass some of the tests or samples, such as low-quality sputum which cannot be analyzed, or when a physician fails to record FEV1 of the particular patient after getting some medicine. Researchers'situations or equipment failures might then play a role in missing values.

As a result of the presence of missing values, bias and inefficiency are likely to result from the statistical analysis of datasets. More precisely, from a statistical point of view, working on an incomplete dataset is very unpleasable and there is not a unique answer to deal with this issue. Loss of power, large value of standard error, wide confidence interval, and less efficiency can be the consequences of missing values. From a statistical point of view, it is always challenging to analyze incomplete data, especially when the number of missing values is substantially increased. In order to analyze a dataset that contains missing values, the first step is to explore the reason for missingness in order to determine the strategy to handle missing values. A method for handling missing values will be chosen in accordance with the number and pattern of missing values, their reasons, and possible implications. Consequently, the first challenge in this study was the investigation of missing values and how to handle them within the framework of cluster analysis.

This section presents the various pattern and mechanisms of missingness, and the corresponding methods to deal with missing values. Finally, since the principal part of this thesis is based on multiple imputation, the well-known multiple imputation process based on chained equations is addressed.

3.1.1 Pattern and Mechanism of Missingness

In general, when analyzing incomplete data, it is necessary to consider the nature of missingness. Two concepts frequently used in missing values are pattern and mechanism. How values are missing in a dataset is called the pattern of missingness while the probabilistic definition of the missing value is called mechanism. The patterns of missingness can generally be divided into monotonous and non-monotone patterns.

When missing values happen in the i^{th} position of an object and all of the subsequent values are also missing, this drop-out pattern is called the monotone pattern. When missing values occur in any of the variables in any position, this pattern is named non-monotone pattern. In this thesis, as presented in subsequent chapters, the considered datasets contain a mix of both monotone and non-monotone missingness patterns.

The mechanism of missingness is defined as the probability of missing values given the other values of variables in the dataset. Missing values mechanisms can be classified into three groups: Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Molenberghs and Kenward, 2007)

Let $\mathbf{X}$ be the incomplete dataset which can be partitioned as $(\mathbf{X}_{mis}, \mathbf{X}_{obs})$ where $\mathbf{X}_{mis}$ is the missing part and $\mathbf{X}_{obs}$ is the observed part. Let L be defined as a random indicator variable for univariate case and random indicator matrix for multivariate case which distinguish missing and observed values for $\mathbf{X}$. If the value is observed for i^{th} object in the j^{th} variable, $L_{ij} = 1$ otherwise, $L_{ij} = 0$. The dependence of L on the variables in the dataset defines the mechanism of missingness. Let define $p(L|\mathbf{X}, \phi)$ as a statistical probability for missing value, and ϕ is the parameter for missing value. In the following, the different mechanisms of missingness are introduced.

Missing completely at random (MCAR)

MCAR is the most common group of missing values. Missing values are involved in this group when the probability of each entry to be missing is independent of the other values in the complete case, $\mathbf{X}$.

$$p(L|\mathbf{X}, \phi) = p(L, \phi).$$

As an example, it is possible that some participants had missing laboratory values due to improper processing of laboratory samples. As a result, under MCAR, missing data depends neither on the observed nor unobserved value.

Missing at random (MAR)

MAR is the second mechanism for expressing missingness that is a special case of MCAR. If the probability of a value to be missing possibly depends on observed data, $\mathbf{X}_{obs}$, but is independent of unobserved data, $\mathbf{X}_{mis}$, this kind of the missing value is

called missing at random. Therefore,

$$p(L|\mathbf{X},\phi) = p(L|\mathbf{X}_{obs},\phi).$$

It is what we commonly refer to as "random". This means that the probability of missingness depends only on observed values in **X**, not on any unobserved values in **X**. A simple example of MAR would be when patients over a particular age refuse to respond to a survey question, and age is an observed covariate.

Missing not at random (MNAR)

MNAR is the third mechanism of missingness. This mechanism of missingness appears if the probability of a value to be missing depends on unobserved data. Therefore, the probability of a value to be missing is dependent on $\mathbf{X}_{mis}$ or some unobserved covariates. As an example of MNAR, objects with incomes above a certain point in the survey might refuse to report such incomes. The missingness depends on the unobserved response, and income (Molenberghs and Kenward, 2007).

3.1.2 Dealing with missing values

In literature, there are several methods for handling missing values. We divide these methods into two simple and advanced groups. Listwise deletion, available case analysis, single imputation, the indicator method, and weighting are contained in the simple approach. While the advanced approaches are likelihood-based methods, posterior-based approaches, and multiple imputation.

Complete Case

One of the methods for overcoming missing values is the complete case method also known as listwise deletion. In this method, objects who contain at least one missing value for one variable are removed, and the remaining complete objects are considered for the statistical analysis. The complete case approach presents several disadvantages such as waste of data, biased and inefficient results. Complete case study was recommended when only a few values are missing completely at random. However, all of the statistical software used this approach as a standard method in the presence of missingness.

Single Imputation

In this approach, each missing value is replaced by a plausible value and then the

imputed dataset is analyzed. Depending on the type of the variables, the simplest and fastest solution is replacing missing values with mean, median, or mode of the observed data for each variable. Another more appropriate method which avoids underestimation of the underlying variance is known as the conditional mean imputation. In this single imputation method, each missing value is imputed by the expected value conditional on observed values. While simple, the single imputation approaches consider the replaced value as the true value and then ignores the fact that imputation method can not provide the exact value. In other word, single imputation approaches do not reflect the uncertainty about the missing values.

Multiple Imputation

Multiple imputation is an approach for filling in missing values with more than one plausible value to account for uncertainty about the prediction of the missing values. In medical research, when our aim is regression estimation, multiple imputation is a popular and very flexible technique for handling missing values. In multiple imputation, each missing value is replaced with a set of $m(>1)$ independent values to generate m separate complete datasets. This approach incorporates uncertainty of the missing data that cannot be achieved with single imputation ($m = 1$). After generating m separate complete imputed datasets, based on Rubin's rules, each of the imputed datasets has to be analyzed individually and in the final step, results must be combined (Rubin, 2004).

In this method, missing value is imputed by drawing from the posterior predictive distribution of Bayesian model. Let $\mathbf{X}$ denote an incomplete data matrix and M the total number of imputed datasets by multiple imputation method. Missing values for m^{th} ($m = 1,2,\cdots,M$) imputed dataset is drawn as:

$$\mathbf{X}_{mis}^{m} \sim p(\mathbf{X}_{mis}|\mathbf{X}_{obs}), \tag{3.1}$$

with

$$p(\mathbf{X}_{mis}|\mathbf{X}_{obs}) = \int p(\mathbf{X}_{mis}|\mathbf{X}_{obs},\theta)\, p(\theta|\mathbf{X}_{obs})\, d\theta, \tag{3.2}$$

where $\mathbf{X}_{mis}$ and $\mathbf{X}_{obs}$ are missing and observed parts of $\mathbf{X}$, respectively. The imputed values are drawn from the joint posterior distribution of missing data given observed data, equation 3.1. In equation 3.2, it is difficult to figure out predictive distribution as long as the requirement of integration over the model parameters θ. This problem is solved in the univariate case by iteratively drawing a sequence of values of the missing data and parameters until convergence. The data is drawn inde-

pendently m times from m approximate posterior distribution involving m estimates for $\theta^{*(1)}, \theta^{*(2)}, \cdots, \theta^{*(M)}$ from $p(\theta|\mathbf{X}_{obs})$ which are used in the conditional distribution $p(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \theta^{*(m)})$ to draw M imputation. However, in nonlinear multivariate data, building one model for the joint distribution of the variable is difficult. In this situation, one of the solutions is multivariate imputation by chained equation (MICE) which is introduced in the following.

3.1.3 Multivariate Imputation by chained equations (MICE)

In a dataset with more than one variable containing missing values, a conditional model is defined for the missing value for each incomplete variable given a set of other variables. MICE divides the p-dimensional problems into p one-dimensional problems. Then, it draws values for the parameters of each incomplete variable and imputes it from the corresponding conditional model. This process iterates through the other incomplete variables. By using this procedure, complex problems like variables with different types, nonlinear models, or interaction between variables and circular dependence can be easily addressed when considering iteratively sampling from conditional distribution instead the algorithm assumes multivariate joint modeling. Conditional models are utilized in MICE without the need for a multivariate model for the whole dataset.

As a reminder, let $\mathbf{X}$ be the incomplete data $N \times P$ matrix with N objects and p variables. $\mathbf{X}$ can be divided into missing and observed parts, $\mathbf{X}_{mis}$ and $\mathbf{X}_{obs}$, respectively. Then, $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$ denotes the joint multivariate posterior which is completely specified by θ, a p-dimensional vector of unknown parameters. One of the main aims of MICE is to obtain the posterior distribution of θ through chained equations which take sampling iteratively from conditional distributions. However, if all p variables contain missing values, MICE starts with a simple draw from observed marginal distributions. Iteration of the chained equation is summarized as follows;

1. Draw $\theta_j^{*(t)}$ from the posterior probability of θ_j; $\theta_j^{*(t)} \sim p\left(\theta_j | x_1^{obs}, x_2^{t-1}, \cdots, x_p^{t-1}\right)$ $j = 1, 2, \cdots, p$;
2. Draw $x_j^{*(t)}$ from the conditional probability of missing values given observed data and estimated parameters θ_j^*; $x_j^{*(t)} \sim p\left(x_j^{mis} | x_1^{obs}, x_2^{t-1}, \cdots, x_p^{t-1}, \theta_1^{*(t)}\right)$ $j = 1, 2, \cdots, p$.

where $x_j^{(t)} = \left(x_j^{obs}, x_j^{*(t)}\right)$ is the j^{th} imputed variable at iteration t and $\theta_1, \theta_2, \cdots, \theta_p$ are the components of θ. When the algorithm converged, the above two steps provide

a draw of θ^* from its posterior which can be used to draw values $\mathbf{X}^*$ to impute $\mathbf{X}_{mis}$. The algorithm has been shown to converge quickly (10 iterations might be enough and throughout this thesis, 10 iterations have been considered) since $x_j^{*(t)}$ is entered by previous imputations $x_j^{*(t-1)}$ through their iteration with other variables. This procedure is repeated M times to generate M imputed datasets.

There are many different techniques for performing multiple imputation, however, the application of predictive mean matching after regression switching (MICE-PMM) appears to be the best method when >10% of data are missing (van Buuren, 2018).

Rubin's Combining rules

Let Q be the parameter of interest, for example, mean or regression coefficients. When analyzing the M imputed datasets, the M estimates $\hat{Q}$ are obtained and need to be combined. The estimate for multiple imputation is simply calculated as the average of the M imputed data estimates,

$$\bar{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}^{(m)},$$

where $\hat{Q}^{(m)}$ is the estimate derived from the m^{th} imputation. In order to calculate the associated standard error, between imputation variance and within imputation variance must be combined. In this case, the variance between imputations is presented by B;

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}^{(m)} - \bar{Q}\right)^2$$

while the within imputation variance is denoted by $\bar{U}$

$$\bar{U} = \frac{1}{M}\sum_{m=1}^{M}U^{(m)},$$

where $U^{(m)}$ is the estimated variance of $\hat{Q}^{(m)}$. Finally, the total variance is equal to

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B,$$

where, $\left(1+\frac{1}{m}\right)B$ estimates the increase in variance because of the presence of missing data and $\bar{U}$ calculates the variance if the data were complete.

This process of inference is very straightforward when it is applied to a population parameter. However, in the context of cluster analysis, the situation is differ-

ent. Indeed, within cluster framework, the objective is to classify each object into homogenous clusters according to the values of the variables, which implies that all inferences are directed at the objects level rather than a population measure. Furthermore, a measure of uncertainty related to cluster assignment is not defined during the clustering process, which makes it unclear how to account for the uncertainty due to imputations in the final results (Rubin, 2004). A major challenge was missing values, which was handled using multiple imputation. Multidimensional datasets and extensive variables in clustering are next on the list of challenges. A detailed look at this challenge will follow in the next section.

3.2 Dimension reduction

In most practical classification studies, a large number of variables are usually considered. Dimension reduction and cluster analysis are two of the most widely used methods which are often come together and have a past history in data analysis. In cluster analysis, when users consider a large number of variables that may contain redundant information with noisy variables, the true structure of classification may be covered up. Datasets with a high number of variables make also interpreting and visualizing the data a challenging task. These challenges are even greater when some of these variables are considered irrelevant or appear to have a small impact on the data structure. In addition, the multidimensional space makes clustering the data more complex and dimension reduction is the only way to proceed with cluster analysis.

There are various methodologies in the literature that aim to reduce the dimension of variables. The methods of dimension reduction generally reduce the number of variables through either selecting important and key variables or the construction of new components using combinations of the original variables. A key aspect of cluster analysis is identifying clusters for all objects, and providing a proper name to each cluster; therefore, regardless of dimension reduction, the original variables must be reviewed when the analysis is finished. As a result, this combination of reducing and clustering data has practical applications in many scenarios and is inevitable.

It has been suggested for cluster analyses that considering variables containing identical information should be avoided. However, there is no unifying statement

between the two methods of dimension reduction, namely selecting variables by expert opinion, or reducing variables by creating new components. Different criteria for variable selection have been developed in cluster analysis. A popular variable selection method is described in the following section that has been suggested by Basagaña et al. (2013), in the study of cluster analysis on incomplete datasets. A common method for variable reduction, known as PCA, is also detailed hereafter.

3.2.1 Variable selection

Selection of variables involves simply excluding some variables and including others without any modification in them. Cluster identification is directly and significantly affected by the outcomes of variable selection. The variables should then be selected in a way that will result in a better cluster output with a high classification rate. It is essential that only relevant variables are kept when performing the variable selection. One of the most simplistic and naive methods of selecting variables is based on the opinion of the researcher and the purpose of the study which defines the relevant variables. The alternative method is to use some criteria to determine key variables. The general concept of this kind of method is to perform an optimization task with regard to a particular objective function, which is then solved using global optimization heuristics. In cluster analysis, several different variable selection methods are available. Some of those methods are described in the following.

Unsupervised variable selection

This method identifies the most important variables and the number of clusters at the same time. In this method, cluster analysis is conducted using different numbers of clusters and different numbers of variables. Final results are then ranked using the *CritCF* criterion to determine which combination of clusters and variables has the highest *CritCF* which reflects high classification. The *CritCF* is defined as

$$CritCF = \left[\left(\frac{2p}{2p+1} \right) \left(\frac{1}{1+\frac{SSW}{SSB}} \right) \right]^{\frac{\log(k+1)+1}{\log(p+1)+1}},$$

where p is the number of considered variables, k is the number of clusters, SSW and SSB are the sum of square within and between clusters, respectively (Section 2.3.1). It is recommended that *CritCF* value be as high as possible and a higher *CritCF* value is preferred to select key variables as well as the number of clusters.

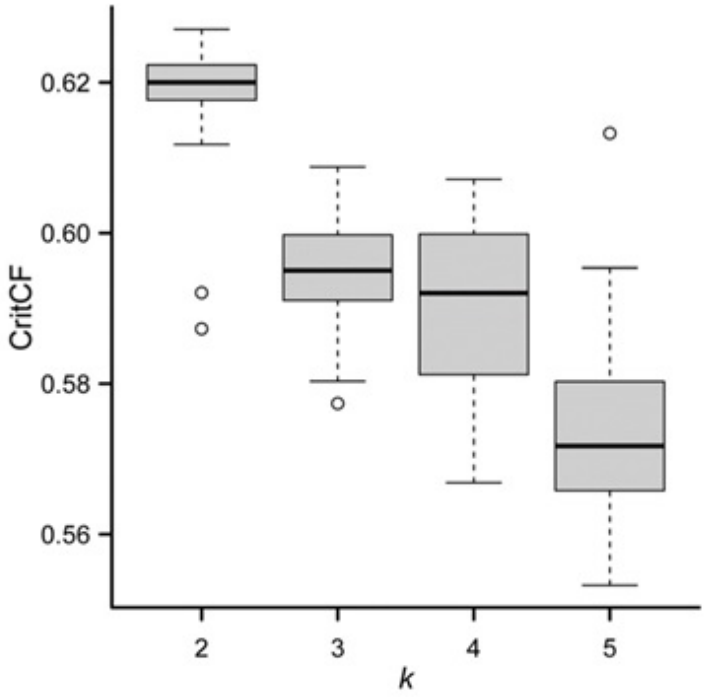

Figure 3.1: Box plots of *CritCF* by the number of clusters (k), PAC-COPD Study, Spain, 2004 – 2008 [Figure from Basagaña et al. (2013)].

In order to find the key variables and the number of clusters, backward sequential selection could be employed. In this strategy, the number of clusters is fixed a priori as, $k = 2,3,\cdots,k_{\max}$, where $k_{\max}$ is the maximum number of clusters and definitely lower than the number of objects. The algorithm starts with two groups of variables, a selected group, and a removed group. In the first step, the selected group contains all variables, while the removed group is empty. As part of each step, the K-means (Section 2.2.1) method is applied by excluding one variable from the selected group. The variable that provides a lower value of *CritCF* when excluded from the selected group is moved to the removed group. Iteration continues until no improvement in *CritCF* is observed. The final selection is chosen based on a combination of cluster numbers and variables that give a higher *CritCF*.

As an illustration, consider Figure 3.1 issued from (Basagaña et al., 2013) which reports *CritCF* calculated for four different numbers of clusters. Since the highest *CritCF* value was obtained for $k = 2$, two clusters were determined for the best homogenous group in the study, and those variables that provided the highest *CritCF* value in the selected group were identified as important and influenced variables for clustering.

Basagaña et al. (2013) have noted that it is possible to design a search strategy based on this criterion to identify both the optimal number of clusters and the final variables to include in the analysis simultaneously. However, there is no evidence to show that this search strategy will be exhaustive, and it is not certain if the number of clusters and variables will reach the global optimum.

3.2.2 Variable reduction

In a multidimensional dataset, variable reduction is a crucial step for accelerating cluster analysis without sacrificing the power of the original variables. Through variable reduction, variables are transformed into fewer dimensions. In general, this method considers all linear combinations of the variables and then constructs new uncorrelated variables (commonly known as principal components, or PCs), and eventually, a reduced number of the components are chosen that still effectively reflected a majority of the information from the original dataset. The number of the selected principal components is usually much smaller than the number of original variables. The number of the components is determined by considering how much information can be preserved while combining the variables of the original dataset. In addition, dimension reduction methods are also effective when the variables of the dataset present collinearity, i.e. some or all of the variables are correlated. The principal components are uncorrelated, so the redundant information is eliminated.

In order to conduct a comprehensive cluster analysis, the datasets in this thesis included all available recorded variables and are composed of both quantitative and qualitative variables. In recent years, the variable reduction technique focused only on quantitative variables, principal component analysis (PCA), or only on qualitative, corresponding analysis (CA). To deal with mixed types of variables, factor analysis of mixed data (FAMD) is used for datasets containing both quantitative and qualitative variables. Using FAMD, quantitative variables are analyzed using PCA, and qualitative variables are analyzed using CA.

For simplicity, only PCA is explained here but this method can be easily generalized to CA and FAMD. The first mathematical step of the PCA is to determine the eigenvalues, and their corresponding eigenvectors, from the covariance matrix of the dataset. The eigenvalues are derived by solving the following equation (Poole, 2006).

$$\det(\Sigma - \lambda \mathbf{I}) = 0,$$

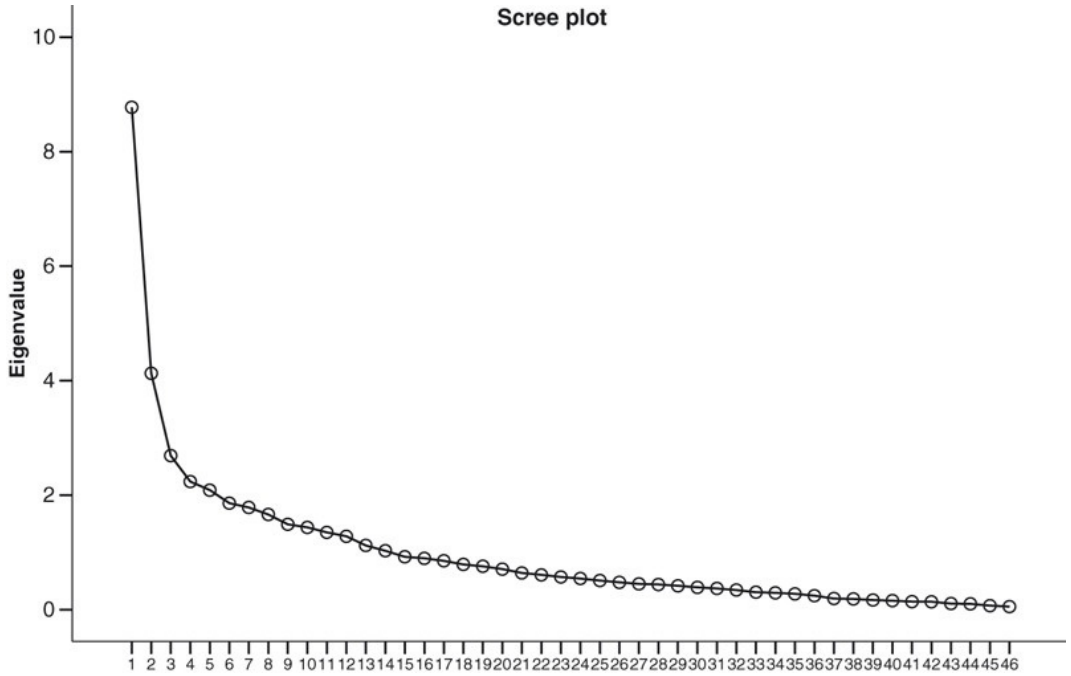

Figure 3.2: An example of a scree plot.

where det is the abbreviation of determinant, Σ is the variance-covariance matrix of the dataset and $\mathbf{I}_p$ is a $p \times p$ identity matrix. Once the eigenvalues are calculated, the corresponding eigenvectors can be derived(Aldrich, 2006). The principal components, i.e. linear combinations of the variables from the original dataset are the results of the application of the eigenvectors to the original dataset.

The derived eigenvalues are used to calculate how much variability of the original dataset can be explained by each component. To do this, each eigenvalue divides by the total sum of the components'eigenvalues. In order to calculate the cumulative sum of the explained variance, it needs to add the prior percentages and the percentage of the component that is being calculated. This cumulative percentage of variance explained is an important criterion used to determine the number of components to consider. The scree plot is a line plot of the eigenvalues of principal components, and it is a widely used method to determine the number of components to consider.

In scree plot, eigenvalues are indicated on the y-axis and number of components are demonstrated on the x-axis. Figure 3.2 illustrates an example of this scree plot. Scree plots typically follow a similar pattern, beginning with a high point on the left, declining relatively rapidly, and then flattening out at the endpoints. In general, the

first component explains most of the variability, the next few components explain a small fraction of the overall variability, and the last few components explain a small fraction of the overall variance. Scree plots operate in the same way as elbow plots (Section 2.2.1), they look like curves and select all components just before the elbow.

In this thesis, we select the number of components that explain at least 90% of the variance. In other words, the first component which its cumulative sum of explained variance is greater than 90% is assumed as an index and this component and all the previous ones are considered for the cluster analysis. It should be noted that PCA does not apply when missing values are present. Therefore, as multiple imputation will be used in this thesis to handle missing values, the components to consider will be determined after PCA is performed on the imputed datasets.

3.3 Consensus Clustering

When multiple imputation is applied for handling missing values, based on Rubin's rules, each of the imputed datasets has to be analyzed individually and in the final step, results must be combined. Therefore, after application of cluster analysis on each imputed dataset, one clustering final result must be achieved. In this case, m individual clustering results are obtained and create an ensemble matrix, R, with N rows for objects, such that, each object is represented as an m-dimensional categorical vector $\{\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_N\}$ and M columns as outputs of clustering,$\Pi = \{\Pi_1, \Pi_2, \cdots, \Pi_M\}$. The m^{th} element of vector $\{\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_N\}$, $c_{i,m}$, represents the cluster assignment of i^{th} object ($i = 1, 2, \cdots, N$) in the m^{th} imputed dataset ($1 \leq m \leq M$) and the result would be an integer value between 1 and k.

Consequently, in the last step of Rubin's process, a method is required to combine the M different clustering results to get a final solution. In the literature, consensus clustering refers to the situation where several clustering methods were applied to a particular dataset to find a single clustering (Topchy et al., 2004; Li et al., 2010). However, here, consensus clustering refers to applying the same clustering method to m different imputed datasets.

In this process, the main problem with clustering based on one specific method on m different imputed datasets is label correspondence throughout the ensemble matrix. For example, clustering on the first imputed dataset may classify it into two

clusters, such that the first twenty percent of data are labeled as cluster 1, and the rest are classified as cluster 2. Whereas in another imputed dataset, the same method may classify into two clusters where the first twenty percent of data would be labeled as the second cluster and the rest as the first cluster. Although the classifications of both imputed datasets are the same, the cluster labelings are different. The cluster labeling is determined by its label assignment process and there are no specific and clear rules for how methods assign labels to determined clusters. Therefore, a relabeling method is necessary to combine the results from imputed datasets. As a general suggestion, the first imputed dataset is taken as a reference and other imputed datasets'clustering results should be harmonized accordingly.

In the present section, two well-known methods for consensus clustering were summarized. Finally, our proposed method which was inspired by model-based clustering on mixture multivariate multinomial distribution is introduced.

3.3.1 Majority Vote

Basagaña et al. (2013) introduced a simple method for consensus clustering. In the beginning, the labeling issue needs to be resolved and harmonized by considering the cluster label from the first imputed dataset as a reference. Then, all possible permutations of labels are considered for other clustering results, and the permutation with the highest agreement with the reference one is selected. Finally, for consensus clustering, the object is assigned to the cluster that is the most frequently observed among the M different clustering results for the object.

$$C_1 = mod_m\left(c_{i,m}\right) i = 1,2,\cdots,N; m = 1,2,\cdots,M. \tag{3.3}$$

However, Majority Vote is restricted to cluster analysis with a fixed number of clusters and predefined important variables for all imputed datasets. This method does not allow to handle clustering with different numbers of clusters for each imputed data.

3.3.2 Co-Membership Method

The second method is based on Co-Membership matrix (Gordon, 1999). Gordon and Vichi (2001) extended Gordon's idea by proposing a method for obtaining a final result from the clustering results of the same dataset derived when using several methods or one method on different dataset. By using this method, consensus clustering was driven by the minimization of a criterion function that measures how sim-

ilar or different consensus candidates are from the ensemble (the method known as the optimization approach to consensus clustering). Bruckers et al. (2018) applied this method for consensus clustering in multiple imputation. Co-Membership matrix is used to represent the similarity between each pair of rows in the ensemble matrix. To determine the objects in the same clusters, $cm_{ij} = 1$ if two objects i and j are in the same cluster, and 0 otherwise. One way to calculation of Co-Membership matrix is,

$$CM(F) = FF^T,$$

where F is Membership matrix which $f_{ik} = 1$ if the object i belong to cluster k, $(k = 1, 2, \cdots, K)$, and 0 otherwise. In Membership matrix, the assigned label for each cluster is not unique. The cluster labels can be modified without changing the clustering results for objects. Therefore, it is necessary to perform a suitable permutation. In terms of the Membership matrix F, arbitrary permuting of this matrix is shown as $F\Pi$, where Π is a suitable permutation.

If we consider $\Phi = (\Phi_1, \Phi_2, \cdots, \Phi_M)$ elements of the m ensemble clustering using Co-Membership matrix and d is a dissimilarity measure, consensus solution is derived from

$$\sum_{m=1}^{M} w_m d\,(C, \Phi_m)^p,$$

where C is the final consensus result where minimizing weighted average dissimilarity power of order p. If we consider similarity measure as d, then when $p = 1$, is calculated from (weighted) median or medoid clustering while $p = 2$, we have least-square consensus clustering. In this thesis, the final consensus clustering result was obtained from least-square fit of a set of Co-Membership on m different clustering results (see, e.g., Gordon (1999); Gordon and Vichi (2001); Hornik (2005); Bruckers et al. (2018)).

3.3.3 Mixture Multivariate Multinomial Model (4M)

Topchy et al. (2004) proposed a method to combine many individual clustering results to provide an improved overall clustering result of the given data. They proposed an expectation-maximization (EM) algorithm with a finite mixture of multinomial distributions for consensus clustering. However, inspired by this idea, 4M is proposed in this thesis, as a novel method for consensus clustering. 4M assumes that the ensemble matrix (R) follows mixture multivariate multinomial model. Then, maximum likelihood was solved using EM algorithm to find the best consensus clus-

tering on ensemble matrix. While Topchy et al. (2004) assumed that the number of clusters is predetermined, this challenge is here solved by selecting the best fitting of mixture as the number of clusters. The number of mixtures is estimated by an approximation of the BIC (Kass and Raftery, 1995; Schwarz, 1978).

Before starting the procedure, we still have to harmonize the labelings. The method of Basagaña was used to define clustering the first imputed dataset as a reference and select the permutation of the other clustering result of imputed datasets that has a high agreement with the reference. The clustering labels in ensemble matrix, R, was assumed nominal values and matrix follows mixture multivariate multinomial distribution. So, in this study, the final result for consensus clustering is found as a solution of maximum likelihood on mixture multivariate multinomial model (4M) using EM algorithm.

It was proposed that ensemble matrix follows mixture of multivariate multinomial distribution

$$p\left(\mathbf{C}_i|\theta\right) = \sum_{g=1}^{G} \Pi_g p_g\left(\mathbf{C}_i|\theta_g\right),$$

where g is the number of mixtures (or clusters) $\left(1 \leq g \leq G\right)$ and define the final number of consensus cluster. The mixing coefficients Π_g correspond to the prior probabilities of the clusters. Since the clustering result for m imputed datasets are independent, all the values, R are assumed to be independent and identically distributed. Since the variables C_i, takes nominal values from a set of cluster labels, it makes sense to view them as being the outcomes of a multinomial distribution. So, $p_g\left(\mathbf{C}_i|\theta_g\right)$ follows multivariate multinomial distribution for i^{th} objects.

$$p_g\left(\mathbf{C}_i|\theta_g\right) = \prod_{m=1}^{M} p_g^{(m)}\left(c_{im}|\theta_g^{(m)}\right),$$

and $p_g^{(m)}\left(c_{im}|\theta_g^{(m)}\right)$ follows multinomial distribution for i^{th} objects in m^{th} imputation

$$p_g^{(m)}\left(c_{im}|\theta_g^{(m)}\right) = \prod_{k=1}^{K(m)} \left[\upsilon_{mg}(k)\right]^{\delta(x,k)},$$

where $\upsilon_{mg}(k)$ presents the probability of the clustering labels such that sum up to

one,

$$\sum_{k=1}^{K(m)} \upsilon_{mg}(k) = 1,$$

and $K(m)$ shows the labels in m^{th} imputation. For example, if m^{th} imputed dataset is classified to two clusters and we labeled the 0 and 1, then probability for m^{th} clustering can be simplified as

$$p_g^{(m)}\left(c_{im}|\theta_g^{(m)}\right) = \upsilon_{mg}^{c_{im}}\left(1-c_{im}\right)^{1-c_{im}}.$$

The log likelihood function for the parameters $\Theta = \{\pi_1, \cdots, \pi_G, \theta_1, \cdots, \theta_G\}$ given the Ensemble matrix $\mathbf{X}$ is:

$$\begin{aligned}\log L(\Theta|\mathbf{R}) &= \log\prod_{i=1}^{N} p(\mathbf{C}_i, \mathbf{z}_i|\Theta) = \log\prod_{i=1}^{N}\prod_{g=1}^{G}\left[\Pi_g p_g\left(\mathbf{C}_i|\theta_g\right)\right]^{z_{ig}} \\ &= \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\left[\pi_g p_g\left(\mathbf{C}_i|\theta_g\right)\right],\end{aligned}$$

$$\Theta^* = \arg\max_{\Theta}\ \log L(\Theta|\mathbf{R}).$$

There is no closed-form solution to the maximum likelihood problem when all parameters $\Theta = \{\pi_1, \cdots, \pi_G, \theta_1, \cdots, \theta_G\}$ are unknown. Therefore, the likelihood function can be optimized using the EM algorithm. we have to define an auxiliary function

$$\begin{aligned}Q(\Theta, \Theta') &= \sum_{z}\log\left(p(\mathbf{R}, \mathbf{z}|\Theta)\right) p(\mathbf{z}|\mathbf{R}, \Theta') \\ &= \sum_{z}\sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\left[\pi_g p_g\left(\mathbf{C}_i|\theta_g\right)\right] p(\mathbf{z}|\mathbf{R}, \Theta') \\ &= \sum_{i=1}^{N}\sum_{g=1}^{G} E\left[z_{ig}\right]\log\left[\pi_g p_g\left(\mathbf{C}_i|\theta_g\right)\right].\end{aligned}$$

The E-step computes the expected values of the hidden variables $E\left[z_{ig}\right]$

$$E\left[z_{ig}\right] = \frac{\pi'_g \prod_{m=1}^{M}\prod_{k=1}^{K(m)}\left(\upsilon_{mg}(k)\right)^{\delta(x_{im},k)}}{\sum_{g=1}^{G}\pi'_g \prod_{m=1}^{M}\prod_{k=1}^{K(m)}\left(\upsilon_{mg}(k)\right)^{\delta(x_{im},k)}}.$$

The M-step maximizes the likelihood by computing new best parameters esti-

mates:

$$\pi_g = \frac{\sum_{1=1}^{N} E\left[z_{ig}\right]}{\sum_{i=1}^{N}\sum_{g=1}^{G} E\left[z_{ig}\right]}.$$

$$\upsilon_{mg}(k) = \frac{\sum_{1=1}^{N} \delta\left(x_{im}, k\right) E\left[z_{ig}\right]}{\sum_{1=1}^{N}\sum_{k=1}^{K(m)} \delta\left(x_{im}, k\right) E\left[z_{ig}\right]}.$$

A general problem in cluster analysis is determining the number of clusters. In the current study, the number of mixtures is determined by fitting 4M under different numbers of mixtures. Then, the best number of mixtures is selected by BIC. So, the appropriate number of clusters is determined as the final number of mixtures in fitting Mixture Multinomial Multivariate Model on ensemble.

3.4 Conclusion

This chapter provided the necessary methods to overcome the challenges of applying cluster analysis to a large incomplete dataset. Further, the presented methods and proposed method for consensus clustering were discussed. Although, the challenges of cluster analysis were addressed in this chapter, however, these methods should be combined together and set up in a framework in order to achieve the final and adequate clustering result. The proposed framework will be presented in Chapter 4 and the effectiveness of the combination of these methods will be evaluated using simulation studies under several scenarios as well as comparison to other competing frameworks.

CHAPTER 4

Integrate cluster analysis framework using multiple imputation

In the previous two chapters, cluster analysis and the challenges of applying it to datasets with large number of variables and the presence of missing values were introduced. In spite of the fact that the identified methods to handle those challenges are well-known and well documented, it is imperative that they perform well together to provide the most appropriate results when applying multiple imputation on cluster analysis in multidimensional incomplete datasets. In the literature, Basagaña presented a full package of methods to manage this process (Basagaña et al., 2013). The methodology introduced by Bruckers et al. can also be considered as an issue to those challenges (Bruckers, 2014). Therefore, in this chapter, Basagaña's method and Bruckers'method are first introduced in sections 4.2 and 4.3, respectively. Finally, our proposed framework for handling missing values using multiple imputation and multidimensional data in cluster analysis is presented in Section 4.4. In the literature, there is no comprehensive comparison between the existing frameworks on cluster analysis. Therefore, in this section, several scenarios were investigated using simulated datasets with known clustering results under different missingness and

This Chapter is based on

Nekoee Zharaei, H. S., Louis, R., Donneau, A.F., Using multiple imputation for cluster analysis with large incomplete data. (manuscript in under-review)

overlapping rates. These scenarios examine the effectiveness of our proposed frameworks using different competitive methods for each step of the proposed framework.

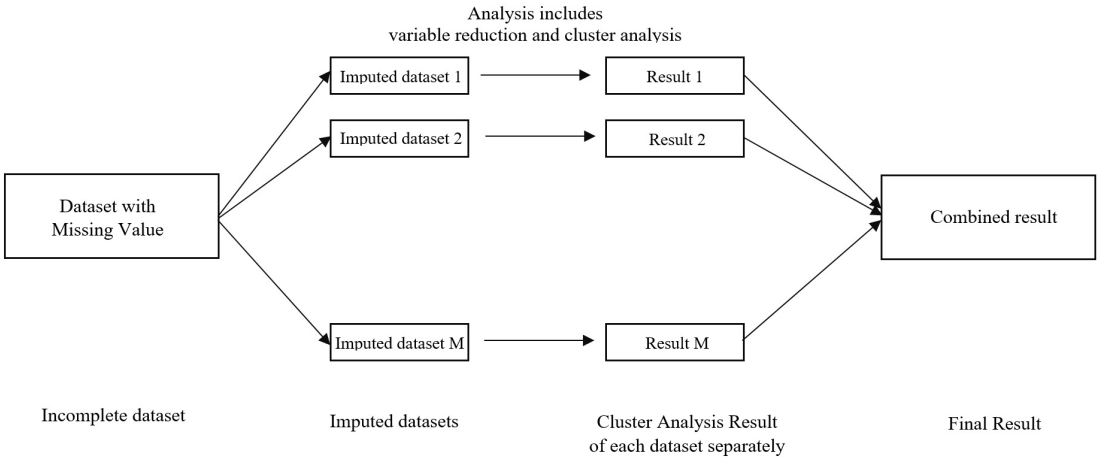

Figure 4.1: The structure of multiple imputation

4.1 Introduction

As explained in Section 3.1.2, when datasets contain missing values for clustering, there are different approaches to handle this challenging issue. A well-known method of dealing with missing values is multiple imputation, which attempts to account for the uncertainty about the prediction of the missing values. When multiple imputation is applied, every missing value is replaced with a set of m independent plausible values and the results are m separate complete datasets. In accordance with Rubin's rules, the imputed datasets must be analyzed separately and the results must be combined at the end. The procedure is illustrated in Figure 4.1. Therefore, using multiple imputation in clustering and obtaining the final result of the clustering requires an integrated framework of linked steps to integrate the treatment of missing values and numerous multidimensional datasets. In general, Figure 4.1 and Table 4.1 demonstrate how such a framework performs for cluster analysis in multidimensional incomplete datasets. The framework algorithm, in Table 4.1, is composed of several competing statistical methods with a lot of decisions to make. Therefore, a summary of the full packages is demonstrated in Table 4.1, and frameworks with different methods and their functions on cluster analysis are provided in the following sections when incomplete datasets are present and various variables are considered.

Table 4.1: The general algorithm of clustering on multidimensional incomplete datasets

Input: Matrix **X** representing a dataset with missing values and multidimensional variables
Step 1. Implement multiple imputation with a predefined number of imputations (m) and iterations (Section 3.1.2), **Step 2.** Utilize one of the existing dimension reduction methods for each of m imputed datasets (Section 3.2), **Step 3.** Apply one method of cluster analysis (Section 2.2) on all m imputed datasets, **Step 4.** After m different clustering results have been obtained in the previous step, one of the consensus clustering methods is used to reach the final result (Section 3.3).
Output: Partition of clustering labels $C = \{C_1, \cdots, C_k\}$

4.2 Basagaña Framework

4.2.1 Definition

Principle framework for cluster analysis based on multiple imputation was introduced by Basagaña et al. (2013). Basagaña et al. were the first researchers who worked on this subject. The framework was started by applying multiple imputation to obtain m imputed datasets. In this method, Basagaña et al. proposed to determine the number of imputations (m) while considering the precision of the object cluster assignment probabilities. For a given object, the corresponding 95% confidence interval for this observed probability, $\hat{p}$, can be computed as

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{m}},$$

where the second part refers to the precision, e. The worst-case scenario occurs when $\hat{p} = 0.5$ and in this case the precision is approximately equal to $e = \frac{1}{\sqrt{m}}$. Therefore, if $m = 100$, the precision would be ± 0.1. This relationship is illustrated in Figure 4.2. When working with large number of variables, the number of imputations, m, will also depend on the computational cost. As a result, 100 imputations were chosen as a common number of imputations for multiple imputation method ($m = 100$).

As presented in Section 3.2.1, Basagaña utilized a backward sequential selection to define the best combinations of variables to consider and the number of clusters to define using *CritCF*. Then, cluster analysis with the determined number of clusters and selected variables were fitted on each m imputed dataset. Finally, each object

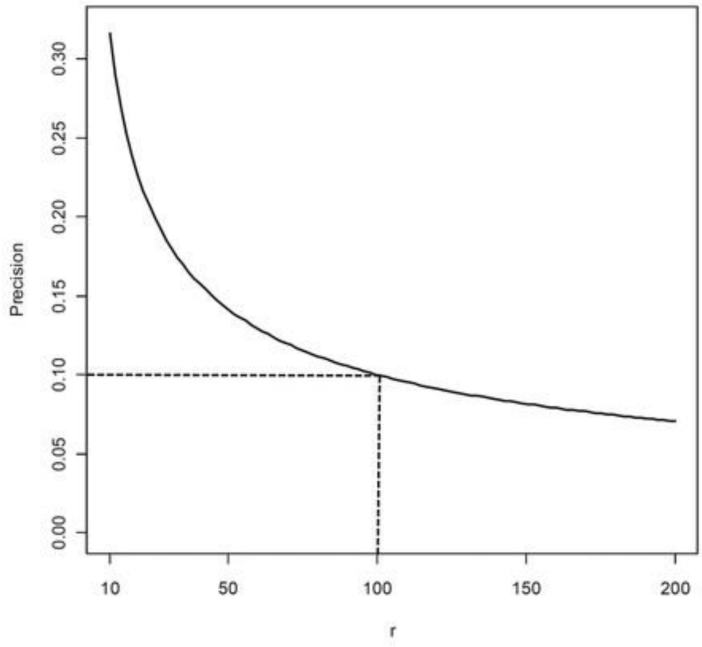

Figure 4.2: Precision of clustering assignment probability for an object as a function of m $(\hat{p} = 0.5)$

was assigned to a cluster using the Majority Vote process (Section 3.3.1). Basagaña et al. proposed a complete R package for multiple imputation in cluster analysis that contains the above steps. Table 4.2 provides a description of the different steps involved in Basagaña's framework.

Table 4.2: Basagaña framework for multiple imputation in cluster analysis

Input: Matrix **X** representing a dataset with missing values and multidimensional variables
Step 1. Implement multiple imputation to obtain $m = 100$ imputed datasets,
Step 2. K-means clustering method was applied on m imputed datasets to determine the available number of clusters ($k = 1, 2, \cdots, k_{max}$),
Step 3. Determine the variables to consider using *CritCF*, $\left(var_{fin}\right)$, and fix the proper number of clusters $k = k_{fin}$,
Step 4. Refit K-means with $k = k_{fin}$, when considering only the selected variables var_{fin} on m imputed datasets,
Step 5. Relabel the clusters and harmonized them by the first result of clustering,
Step 6. Assign the objects into clusters according to the Majority Vote method.
Output: Partition of clustering labels $C = \{C_1, \cdots, C_k\}$

Basagaña et al. illustrated the effectiveness of the framework using a subset of data from the Phenotype and Course of Chronic Obstructive Pulmonary Disease (PAC-COPD) Study (Basagaña et al., 2013). The purpose of this Spanish study (2004-2008) was to identify clinically and epidemiologically relevant subtypes based on clustering methods of COPD. In this study, a comprehensive set of clinical, functional, and biological variables had been considered for a cohort of patients with COPD. A total of 342 participants were recorded by 85 variables. Variables with missing values ranged from 0% to 47.7%, with 68.2% of variables missing less than 5% of values. Only 13.7% of patients presented complete values for the 85 variables. Overall, 5.9% of the values were missing.

According to Basagaña et al. framework, 48 of the original 85 variables were considered as effective in at least one imputed dataset. Indeed, as shown in Figure 4.3 the median number of selected variables was for the first 16 variables, while the third quartile contained 18 variables. The selection of 16 variables corresponds to selecting variables that appear in more than 50% of the imputed datasets. In their illustration, Basagaña et al. chose to continue the framework with the 16 most frequent variables.

After the framework has been completed, two clusters were identified and the results were presented in two formats: raw data and imputed data. In the raw data columns, Table 4.3, the final cluster assignment decided by Majority Vote has been implemented on the raw dataset with missing values. Then, the mean of each of the selected variables was calculated after missing values were excluded from the variables. In the imputed data columns, the final cluster assignment has been applied on m imputed datasets, then the means of the selected variables were calculated for all m imputed datasets. Finally, the median of all the m means was calculated for presentation in Table 4.3.

From this Table, it appears that Cluster 2 exhibited worse symptoms in a variety of areas, including more airflow limitation, higher hyperinflation, more dyspnea, and worse quality of life, compared to patients in cluster 1 whose symptoms were less severe in all domains.

4.2.2 Simulation

Basagaña designed four simple simulation scenarios to evaluate the introduced framework. The simulated datasets were designed for four different missingness

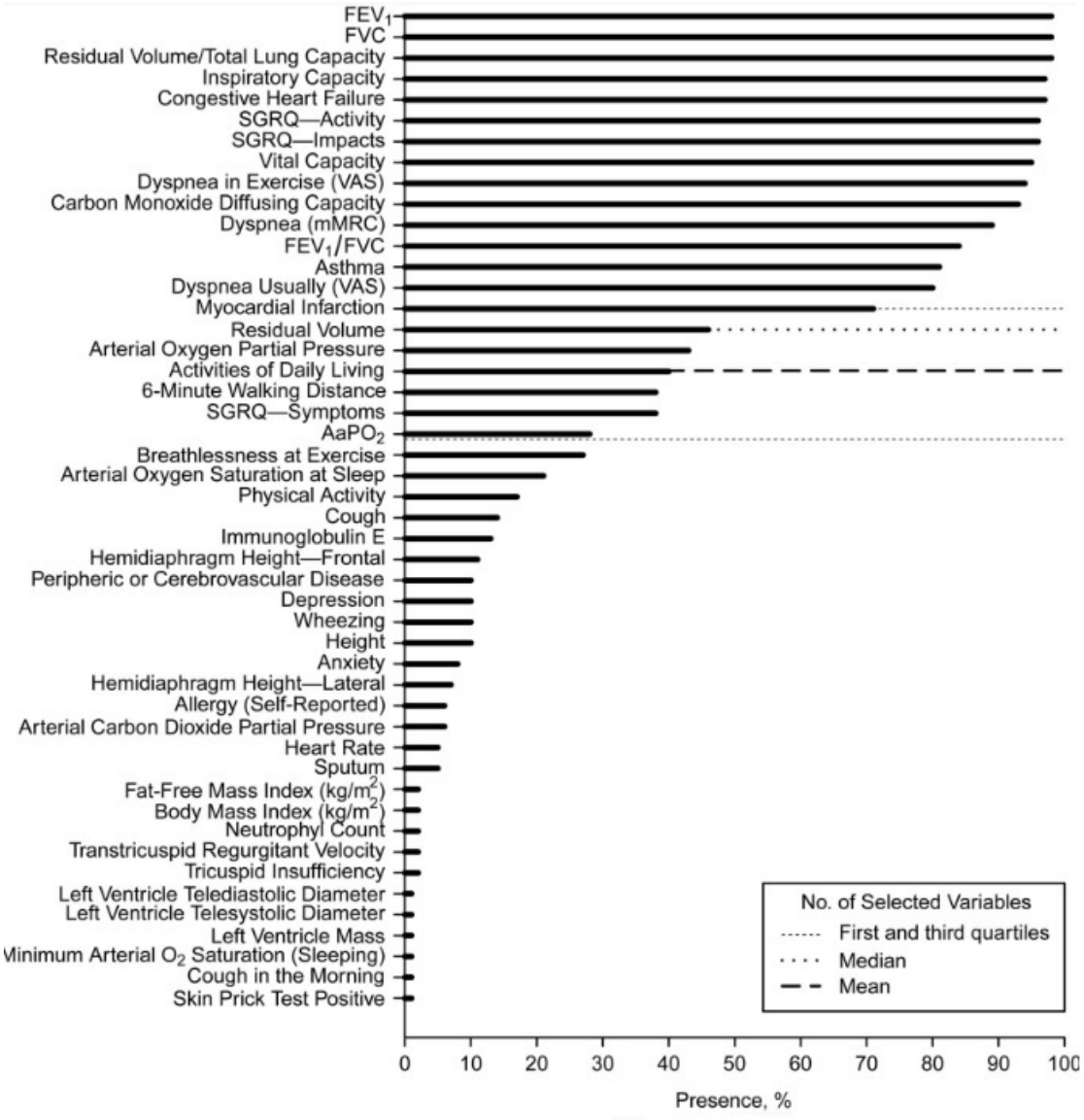

Figure 4.3: Variables that were determined to be important out of 85 variables. (Basagaña et al., 2013)

Table 4.3: Description of the selected variables (mean values) for each cluster, PAC COPD Study, Spain, 2004 – 2008 (Basagaña et al., 2013).

Variable	Raw Data		Imputed Data	
	Cluster 1 (n = 186)	Cluster 2 (n = 156)	Cluster 1 (n = 188)	Cluster 2 (n= 154)
FEV1, % predicted	62.0	41.0	61.7	40.7
FVC, % predicted	79.6	64.4	79.3	64.4
Residual volume/total lung capacity, %	50.6	61.8	50.7	61.8
Inspiratory capacity, % predicted	71.8	51.5	71.3	51.7
Congestive heart failure	4.3	8.4	4.9	8.3
SGRQ—activity (range, 0 – 100)	33.1	64.0	33.9	64.3
SGRQ—impacts (range, 0 – 100)	16.9	37.7	17.4	38.1
Prebronchodilator vital capacity, % predicted	76.9	59.8	76.6	60.0
Dyspnea in exercise (VAS) (range, 0 – 10)	4.1	6.4	4.1	6.5
Carbon monoxide diffusing capacity, % predicted	73.1	55.6	72.1	55.5
Dyspnea mMRC (range, 0 – 5)	2.0	3.3	2.0	3.3
FEV1/FVC, %	58.1	47.8	58.1	47.5
Asthma (Self-reported)	7.1	11.0	7.5	11.3
Dyspnea usually (VAS) (range, 0 – 10)	1.9	4.2	1.9	4.2
Myocardial infarction	8.7	12.9	9.4	12.7
Residual volume, % predicted	140.6	172.6	140.3	172.4

types. Using the original dataset for clustering, 10 variables out of 85 variables in this study were selected containing 5 important variables and 5 variables that were never selected as important variables in the literature. Out of the 342 patients, 298 had complete cases. Using this complete case dataset, they conducted K-means cluster analysis and referred to the result as the "truth" result for clustering when there are no missing values in the dataset. Then, by designing four scenarios, missing values were artificially generated in the complete case dataset and Basagaña's framework was applied to the incomplete dataset to evaluate the efficiency of the framework. Finally, the derived results were compared to the correct results on 100 repetitions. These scenarios were outlined as:

Scenario 1**:** In the first scenario, missing values came from the stochastic model. The probability of missing values in each variable was considered as a function of the other variables in the original dataset. Finally, 24% of the objects had missing values for at least one variable.

Scenario 2**:** Using the same procedure as in scenario 1, the overall percentage of cases with missing data was increased until 70% of data has been completed.

Scenario 3**:** Missing values were created just for the first variable (the first variable

plays the most important role in clustering). The remaining 9 variables had the same structure. The percent of the complete case was set to 80%.

Scenario 4**:** Following the same procedure as Scenario 3, the percentage of missing values in the first variable was increased until 66% of complete cases were obtained.

For the "truth" result, two clusters were determined, and based on the variable selection process in *CritCF*, 5 variables were chosen, which were the 5 most important variables. For scenarios 1 to 4, the average number of selected variables was 3.2, 2.0, 2.7, and 2.9 and in all clustering scenarios, the first variable that was important was selected. Furthermore, there were two clusters in all simulations for the complete case datasets across all scenarios. According to scenario 1.66% of simulations determined two clusters and 34% proposed three clusters. For the other scenarios, the probability of two clusters was higher than three clusters, and finally, two clusters were considered in every simulation for four scenarios.

Then, Basagaña applied Cohen's Kappa coefficient to measure the agreement between the "truth" classification and the calculated classification. The Cohen's Kappa coefficient (κ) is a criterion that is used to measure the degree of agreement between the classification of independent objects. The criterion is defined as follows,

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where, p_0 is the relative observed agreement among clusters, and p_e is the agreement between the result of the simulation and the "truth" cluster as if the result had happened by chance. If all of the objects were classified in correct clusters, $\kappa = 1$, and otherwise, when objects were classified in clusters that were different from correct clusters, $\kappa = 0$. The simple definition is when two clusters are defined, assume the following traditional 2×2 confusion matrix,

Table 4.4: 2×2 confusion table for calculated result for clustering and "truth" cluster

		Result of the clustering	
		Cluster 1	Cluster 2
"truth" cluster	Cluster 1	True positive (TP)	False negative (FN)
	Cluster 2	False positive (FP)	True negative (TN)

where TP and TN mean the clustering result is the same as the truth cluster, and FN indicates objects that should be in cluster 1, classified wrong in cluster 2, and in

contrast, FP indicates objects that should be in cluster 2, classified wrong in cluster 1. In this case, Cohen's Kappa formula is as follows,

$$\kappa = \frac{2\,(TP + TN + FN + FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)}.$$

The highest value for Cohen's Kappa represents a high agreement between the result of clustering and the "truth" cluster and shows the high quality of the proposed framework to assign clusters to the objects.

Comparison of the classification from the original data and that derived from simulation when applying the Basagaña framework revealed that Cohen's Kappa was extremely high (Table 4.5). This result means that sufficient agreement was found even when only 66% of objects were complete and the important variable had a high percentage of missing values.

Table 4.5: Mean and 25^{th}(P25) and 95^{th}(P95) percentages of the Cohen's Kappa obtained over 100 repetitions in each scenario

	Basagaña Framework			Complete case		
Scenario	Mean	P25	P75	Mean	P25	P75
Scenario 1	0.97	0.93	0.99	0.90	0.75	0.99
Scenario 2	0.94	0.92	0.98	0.96	0.93	0.98
Scenario 3	0.95	0.98	0.98	0.84	-0.01	0.98
Scenario 4	0.91	0.88	0.93	0.77	0.78	0.79

4.2.3 Conclusion

The first proposed framework using multiple imputation in cluster analysis was provided by Basagaña et al. This framework is very flexible and well documented. In Basagaña's framework, and in the step of selecting important variables as well as the number of clusters, the authors showed a lot of emphasis on considering uncertainty in multiple imputation. However, the authors believed that no significant difference should be observed between the m imputed datasets. Therefore, it is expected that the majority of imputed datasets are likely to yield the same number of clusters and also the same results for selecting important variables.

They pointed out that, for example, out of 100 imputed datasets, 99% identified the number of clusters as two, while only one dataset identified three clusters. In

addition, of all 85 variables in the dataset, 15 to 18 variables are considered effective variables on 100 imputed datasets, and finally, 16 variables were selected based on *CritCF*. They noted that this occurred since missing data seemed to introduce little uncertainty into this decision and multiple imputation have handled it well. This part of the framework has been extensively discussed, including the choice of a fixed number of clusters and the independence in selecting important variables throughout of imputed dataset.

Furthermore, the Majority Vote method is a simple method that is based on specific assumptions that have performed well if they are held. It is also necessary to note that this framework is time-consuming as it takes into account all possible conditions for the number of clusters as well as for identifying variables and then refitting the cluster analysis for 100 completed datasets.

4.3 Bruckers Framework

4.3.1 Definition

Bruckers (2014) proposed a framework for clustering on high dimensional multivariate longitudinal data with missing values. Despite the fact that Bruckers focused on an incomplete longitudinal dataset, this study addressed the challenges associated with cluster analysis on incomplete datasets which requires variable reduction. Therefore, although the datasets considered in this thesis and Bruckers are different, the overall process is similar. A part of clustering longitudinal data, that had to be accomplished, was dimension reduction on observed times. To achieve this, they considered principal component analysis, which required complete data without missing values. Multiple imputation strategy was chosen in this case to resolve the issue. After applying appropriate multiple imputation and variable reduction methods for the longitudinal data, the same challenge was encountered for clustering on the incomplete dataset. Following that, the model-based clustering technique (2.2.3) was executed for each of the imputed longitudinal datasets and one clustering result was obtained for each of the imputed datasets. To summarize the clustering results for each imputed dataset into a final cluster result, the Co-Membership method (Section 3.3.2) was applied for the consensus clustering part. The steps of their framework for handling cluster analysis on the incomplete dataset were summarized in Table 4.6, regardless of the type of dataset.

Table 4.6: Framework for cluster analysis using multiple imputation on multivariate functional data in Bruckers method

Input: Matrix **X** representing a longitudinal dataset with missing values and multidimensional variables
Step 1. Implement multiple imputation to obtain $m = 10$ imputed datasets, **Step 2.** Apply multivariate functional principal component analysis on m imputed datasets, **Step 3.** Model-based cluster method is applied for functional data with criteria BIC, **Step 4.** Assign the objects into clusters according to the Co-Membership method.
Output: Partition of clustering labels $C = \{C_1, \ldots, C_k\}$

In contrast to Basagaña framework, this study fixed the number of multiple imputations at 10 due to the complexity of longitudinal datasets in this procedure. The performance of Bruckers'framework is illustrated in the following simulation section.

4.3.2 Simulation

For the purpose of demonstrating numerically the validity of the proposed framework (Table 4.6), the authors designed the following model to simulate two clusters of bivariate functional data.

$$\text{Cluster 1: } X_1(t) = 5 + \tfrac{t}{2} + U_2 h_3(t) + U_3 h_2(t) + \sqrt{0.1}\varepsilon(t)$$
$$X_2(t) = -5 + \tfrac{t}{2} + U_1 h_1(t) + U_2 h_2(t) + U_3 h_2(t) + \sqrt{0.5}\varepsilon(t)$$

$$\text{Cluster 2: } X_1(t) = U_3 h_2(t) + \sqrt{10}\varepsilon(t)$$
$$X_2(t) = U_1 h_1(t) + U_3 h_2(t) + \sqrt{0.5}\varepsilon(t)$$

with $U_1 \sim N\left(0.5, \frac{1}{2}\right)$, $U_2 \sim N\left(0, \frac{1}{12}\right)$, $U_3 \sim N\left(0, \frac{2}{3}\right)$, and $\varepsilon(t) \sim N(0,1)$ that are generated independently. Since this study is on longitudinal data, the functions h_1, h_2, h_3 are defined for time $t \in [1,21]$, as $h_1(t) = (6 - [t-11])_+$, $h_2(t) = (6 - [t-7])_+$, $h_3(t) = (6 - [t-15])_+$, where $()_+$ derives the positive part from parentheses. The sample size for the complete simulated datasets was considered 50 observations and the curves were observed in 41 equidistant points for $t = 1, 1.5, \cdots, 21$. The authors considered three percentages for missing values, 10%, 20%, and 30%. For each set, 250 incomplete datasets were simulated and then 10 imputed datasets were generated for each.

The framework follows Table 4.6 and includes multiple imputation on multivariate functional data, functional PCAs as a data reduction technique, model-based clustering on the results of the PCA for functional data, and Co-Membership consensus clustering with a Euclidean distance measure which was executed on simulated datasets (Bruckers et al., 2017). BIC information criteria were used to determine the optimal number of clusters in model-based clustering and scree plot was applied for selecting the number of components in PCA. Due to these conditions, it was not possible for the author to be guaranteed that the simulated data in different models were identical, and the same number of clusters and classifications could not be reached.

The validation of the framework on incomplete datasets was evaluated based on the proportion of correctly classified observations for each simulated dataset. The simulation study results were evidence that the method performs well on longitudinal datasets using the defined percentages for missing data. Most observations are classified correctly into the correct cluster in $72-80\%$ of cases. The results are presented in Figure 4.4. Although the methods used in the framework have a number of sources of uncertainty, noise, errors in incomplete observations, and several uncertainties in the estimated principal component scores and eigenfunctions, the results demonstrated how well it is able to distinguish the cluster structure in the simulated data.

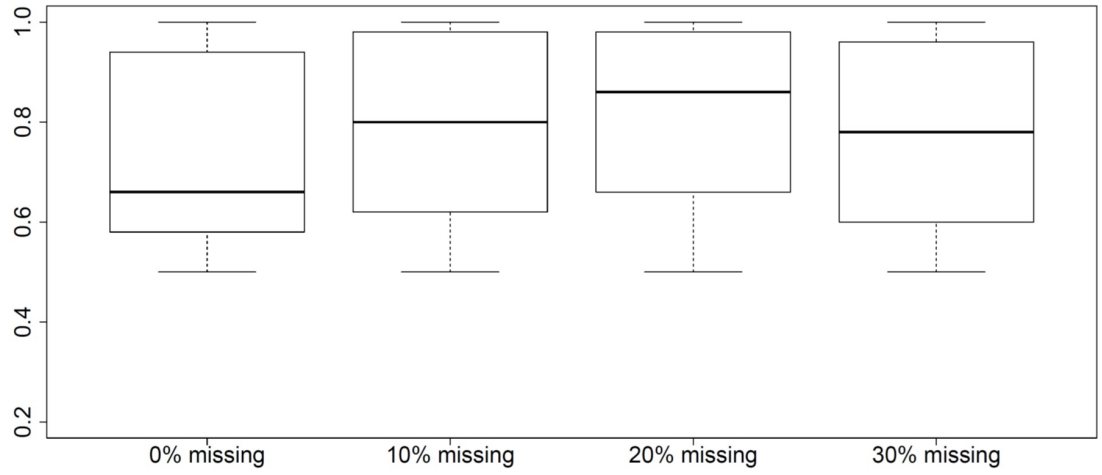

Figure 4.4: Rate of correct classification in simulation study

4.3.3 Conclusion

In this study, Bruckers et al. generated 10 imputed datasets and noted the choice of the number of imputed datasets is still an open topic. Methods that are included in Bruckers'framework and achieve the precision classification become complicated when the number of variables increases. Therefore, this framework cannot be easily implemented and the authors suggested applying the information criterion, *CritCF*, instead of variable reduction so that both the number of clusters and the number of variables are taken into consideration at once. As a result, choosing between variable reduction methods and methods for selecting key variables remains challenging when the number of variables increases. In this study, instead of using K-means, model-based clustering was applied, and there is no evidence of priority between the two approaches.

4.4 Proposed Framework

4.4.1 Definition

In this thesis, we have identified several challenges when apply clustering technic to large variables dataset with missing values. Many studies and numerous methods are present in the literature for each introduced challenge. However, there is no evidence to support the claim that one method is more efficient or effective than another. Therefore, after comprehensively evaluating the benefits and disadvantages of all these methods, we proposed a new framework which combines efficient identified methods to apply clustering technic on large variables dataset with missing values.

As a result, in the first step, multiple imputation was considering as a powerful method in handling missing values, despite its difficulties in cluster analysis. The next step involves reducing the dimensions, and selecting the key variables. This step is always challenging and ambiguous, as there is no common agreement on the most relevant variables. In this thesis, the method of variable reduction was chosen depending on the type of dataset and the criterion for selecting the number of components was determined by ensuring that components explain at least 90% of the variance.

At the cluster analysis step, no evidence in the literature suggests that the methods discussed in Chapter 2 are preferred, so both K-means and hierarchical clustering

were considered in the framework. In the hierarchical clustering, 30 criteria are available for determining the number of clusters. When considering those 30 criteria, the final number of clusters is based on the agreement result. So, the hierarchical clustering method was applied to determine the number of clusters. Then, the objects are assigned to clusters using K-means method. This clustering procedure was applied separately to each imputed dataset. Currently, each of the available methods defined in the literature for consensus clustering is based on a particular condition that limits the components of variable reduction and the number of clusters that can be chosen. Therefore, in the last step of the proposed framework, 4M method (Section 3.3.3) was applied which there are no restrictions for it. Based on the above steps, the proposed framework was developed as shown in Table 4.7.

Table 4.7: Proposed framework for multiple imputation in cluster analysis

Input: Matrix **X** representing a dataset with missing values and multidimensional variables
Step 1. Implement multiple imputation to obtain $m = 100$ imputed datasets (Section 3.1.2),
Step 2. In each imputed dataset, reduce the number of variables by using FAMD when ensuring that the retained components explain at least 90% of the variance (Section 3.2.2),
Step 3. Hierarchical clustering is applied to determine the number of clusters for each imputed dataset (Section 2.2.2),
Step 4. K-means clustering method is applied to each imputed dataset based on its retained principal component (Step 2) and determined number of clusters (Step 3),
Step 5. 4M method is applied to combine the ensemble clusters and achieve the final best cluster result (Section 3.3.3).
Output: Partition of clustering labels $C = \{C_1, \ldots, C_k\}$

This proposed framework needs to be evaluated and compared to parallel frameworks using alternative methods and other complete frameworks (Basagaña's framework, and Bruckers'framework). In the next section, the proposed framework will be evaluated using five different scenarios on a simulated dataset that is inspired by the real dataset. Each scenario was designed to compare competitive methods to find out the best classifications.

4.4.2 Simulation study

The main purpose of the simulation study is to assess the accuracy and efficiency of the new proposed framework for applying cluster analysis correctly to an incom-

plete dataset with large number of variables. To this aim, we focus on assigning objects to the correct cluster and determining the correct number of clusters. We achieved these objectives using generating a dataset that contained a determined fixed number of clusters and the known cluster assignment of object. In our simulations, we attempted to simulate a dataset as similar as possible to real datasets. The simulated datasets with 178 observations and a mixed continuous and categorical variables were generated based on two known clusters.

Mixed data-generated model

More specifically, the main idea of the following simulations was inspired by the *Kamila* package which implements methods for clustering mixed-type data, specifically combinations of continuous and nominal data. In order to simulate mixed dataset with known clustering, Foss and Markatou (2018) considered a matrix with N independent and identically distributed observations of $R = (P + Q)$ dimensional vectors of random variables, first $\mathbf{V}$ columns present P-dimensional vectors of continuous random variables and the next $\mathbf{W}$ columns are Q dimensional vectors of categorical random variables.

In the original program of Kamila, q^{th} element of $\mathbf{W}$ had L_q categorical levels which define by $1, 2, \cdots, L_q$, however, in our simulation studies, dichotomous variables with two levels, $(0, 1)$, were defined for all categorical variables. The vectors $\mathbf{V}$ and $\mathbf{W}$ can be considered dependent and may assume independent within any particular clusters, however, in the following simulations, two vectors are dependent.

Based on membership in the k^{th} cluster, the first P columns, $\mathbf{V}$, following mixture of normal distribution with individual component density function $\varphi_{V,k}\left(v, \mu_k; \Sigma_k\right)$ where k is cluster membership, μ_k and Σ_k denote mean and variance-covariance matrix for k^{th} cluster, respectively. The second Q columns, $\mathbf{W}$, follow finite mixture of multinomial distribution with two levels and individual component probability function

$$f_{W,k}(w, \theta_k) = \prod_{q=1}^{Q} \theta_{kq}^{w} \left(1 - \theta_{kq}\right)^{1-w},$$

where θ_{kq} is the probability vector for k^{th} component of the q^{th} categorical variable. If two groups of continuous and categorical variables will be considered independent, in accordance with membership in the k^{th} cluster, the joint density of $(\mathbf{V}, \mathbf{W})$

will be,

$$f_{V,W,k}\left(v,w,\mu_k;\Sigma_k,\theta_{kq}\right)=\varphi_{V,k}\left(v,\mu_k;\Sigma_k\right)\prod_{q=1}^{Q}\theta_{kq}^{w}\left(1-\theta_{kq}\right)^{1-w}).$$

Finally, the overall density unconditional on cluster membership is

$$f_{V,W}\left(v,w\right)=\sum_{k=1}^{K}\pi_k f_{V,W,k}\left(v,w,\mu_k;\Sigma_k,\theta_{kq}\right),$$

where π_k is a prior probability for k^{th} cluster.

Overlap generating mechanisms

One of usage of this method is that this function simulates all of the steps based on the given membership of k^{th} cluster and this parameter allows to control the degree of cluster separation or overlap separately in both continuous and categorical variables when mixed-type datasets are created. In this method, the overall overlap between two clusters can be defined. This overall overlap was parameterized as the overlap area under their densities.

In this simulation study, the number of continuous and categorical variables was specified based on a real dataset. The simulated datasets contain 66 continuous variables and 20 binary variables. However, for one of the scenarios, the effect of the number of variables and the criteria for dimensions reducing of variables in cluster analysis are evaluated, so, this scenario is run twice with the above condition for 86 mixed variables and also for 56 variables that include 46 continuous variables and 10 binary variables.

For the ability to control separation in the structure of clusters, the degree of overlapping between the two simulated clusters, π_k, was defined as an important parameter in the simulation. Two clusters with zero value, $\pi_k = 0$, of overlapping define a complete separation of clusters and complete overlap happens when the value of overlapping equals one, $\pi_k = 1$. Therefore, several specific degrees of overlapping were considered to evaluate the ability to detect the correct classification; i.e., $\pi_k = 30\%, 45\%, 65\%$.

Missing data generating mechanisms

Finally, for all of the scenarios, 35% of observations were allocated to the first cluster, and the second cluster was contained other simulated observations. In general,

missing values were generated completely at random from uniform distribution with different numbers of missing values; i.e., 10%, 20%, 30%, 45%, and 65%. However, in one of the scenarios, the function of the framework was evaluated in three types of missingness. In that case, missing values for all of the types were generated using *missMethods* package in R studio (Santos et al., 2019).

Evaluation criteria

Two indices were defined for investigating and comparing clustering results among all different scenarios. First of all, it is critical to obtain the correct number of clusters after applying the framework to simulated datasets. Therefore, among the repetitions, the percentage of repetitions that recognize two clusters for the imputed dataset was calculated and considered as one of the criteria to evaluate the well-functioning clustering.

The second criterion is based on the fact that objects should be assigned correctly to the clusters. Therefore, since the reference cluster is known, so, the agreement between the result of the clustering and the reference cluster can be determined. In this case, like in Basagaña, Cohen's Kappa analysis was conducted to evaluate the agreement.

4.4.3 Results

First scenario; missing data

The main purpose of the first scenario was to compare the performance among the complete case approach, the proposed method using multiple imputation, and hierarchical cluster analysis which has the ability to apply to incomplete datasets. Therefore, in this scenario, the necessity of excluding, imputing, or including missing values was evaluated. For this scenario, missing values were generated as being completely at random with percentages of missing values equal to 10%, 20%, 30%, and overlap was fixed at 30%, 45%, and 65%.

In the complete case approach, as explained in Section 3.1.2, all of the observations with at least one missing value were excluded from the simulated datasets, therefore, incomplete datasets were transformed into complete datasets, and therefore, statistical analysis using software like principal component analysis is not a challenge. In this case, after having skipped Step 1 in Table 4.7, the other steps, Step 2 to Step 4, were performed on the completed dataset. In order to calculate the de-

gree of agreement, since objects with at least one missing value were excluded, the Cohen Kappa criterion was calculated on the available observation in both clusters. The final clustering result was reported in Table 4.8 in Complete Case Study columns.

In the second structure which analyze the use of multiple imputation, the process presented in Table 4.7 was applied to the incomplete datasets. The corresponding Cohen's Kappa coefficients between the achieved result and the reference cluster are presented in Table 4.8.

For the last structure, we bypassed the use of multiple imputation by applying directly hierarchical cluster analysis on the distance matrix derived from the incomplete datasets. The Gower's Distance (Section 2.1.2) was calculated for the incomplete dataset, which includes mixed continuous and categorical variables. Then, hierarchical clustering was performed on the distance matrix. The result of the agreement between the assigned cluster and the reference cluster was reported in Table 4.8.

Table 4.8: Kappa Cohen coefficients (95% CI) and percentage of the correct number of clusters derived for First Scenario

			Complete dataset		Multiple Imputation		Hierarchical clustering on incomplete data	
Overlap (%)	Missing (%)	Nr. Cluster	Achieved Clusters (%)	Kappa (95 % CI)	Achieved Clusters (%)	Kappa (95 % CI)	Achieved Clusters (%)	Kappa (95 % CI)
	10%	2	0%	-	100%	0.93(0.88-0.98)	8%	0.78(0.68-0.88)
		≥ 3	99%	0.04(0.05-0.12)	0%	-	92%	0.23(0.15-0.31)
30%	20%	2	0%	-	100%	0.91(0.84-0.97)	11%	0.63(0.54-0.75)
		≥ 3	96%	0.04(0.05-0.13)	0%	-	89%	0.03(0.05-0.11)
	30%	2	0%	-	100%	0.90(0.83-0.97)	14%	0.61(0.52-0.69)
		≥ 3	94%	0.03(0.06-0.13)	0%	-	86%	0.03(0.05-0.11)
	10%	2	0%	-	100%	0.92(0.85-0.98)	0%	-
		≥ 3	91%	0.04(0.06-0.13)	0%	-	100%	0.23(0.16-0.32)
45%	20%	2	0%	-	100%	0.94(0.89-0.99)	0%	-
		≥ 3	99%	0.04(0.05-0.13)	0%	-	100%	0.03(0.05-0.11)
	30%	2	0%	-	100%	0.92(0.85-0.98)	0%	-
		≥ 3	94%	0.04(0.05-0.13)	0%	-	100%	0.21(0.14-0.30)
	10%	2	0%	-	100%	0.94(0.88-0.99)	0%	-
		≥ 3	100%	0.02(0.07-0.11)	0%	-	100%	0.03(0.05-0.11)
65%	20%	2	0%	-	99%	0.91(0.83-0.98)	0%	-
		≥ 3	89%	0.04(0.05-0.13)	1%	0.05(0.03-0.17)	100%	0.03(0.05-0.12)
	30%	2	0%	-	100%	0.92(0.85-0.98)	0%	-
		≥ 3	92%	0.03(0.06-0.12)	0%	-	100%	0.04(0.05-0.12)

In spite of the correct number of clusters being two, Table 4.8 shows that both the complete dataset and hierarchical method identified three clusters or more. The proposed multiple imputation method distinguished the correct number of clusters and classified observations with the high agreement.

Cluster analysis methods cannot be applied directly to incomplete datasets. In the statistical software, the default method for handling missing values is a complete case study. However, the main objective of cluster analysis is to classify the objects into homogeneous clusters, while, the complete case approach excluded the incomplete objects, so, the complete case study is not an applicable method in cluster analysis.

The hierarchical clustering method offers a solution for incomplete datasets and can be applied to distance matrices instead of incomplete datasets. However, in calculating the distance matrix for example using Gower's Distance if two objects cannot be compared due to missing values, the distance matrix value will be zero. As a consequence, the effect of the variable that was not observed will not be considered in the classification of the object. In addition, in presence of missing values, the method of variable reduction could not be utilized. Therefore, Step 2 from Table 4.7 was not applied and hierarchical clustering was here applied to the available data. In addition, when generating simulation datasets, the continuous and categorical variables were considered correlated. Therefore, it is quite possible that these reasons contribute to the fact that, in the hierarchical method, the objects could not be assigned to the correct clusters. Finally, the proposed framework (with multiple imputation) could accurately determine the number of clusters and assign objects to those clusters with greater than 90% agreement.

Second scenario; type of missing value

This scenario was designed to evaluate the performance of the proposed framework under different types of missingness mechanism. According to this aim, after simulation of mixed datasets, different rates of missing values were generated using mechanisms previously described. The proposed framework (Table 4.7) was then applied to the simulated datasets. As *MICE* package provides several methods for multiple imputation based on the type of missingness in the datasets, in Step 1, multiple imputation was applied to each incomplete dataset according to the type of missingness. The results were shown in Table 4.9. It is expected that if the proper multiple imputation method is chosen and 100 imputed datasets are imputed correctly, the results will present high agreement across all types of missing data.

Table 4.9: Kappa Cohen coefficients (95% CI) and percentage of the correct number of clusters derived for Second Scenario

		MCAR		MAR		MNAR	
Overlap (%)	Missing (%)	Achieved Clusters(%)	Kappa (95 % CI)	Achieved Clusters(%)	Kappa (95 % CI)	Achieved Clusters(%)	Kappa (95 % CI)
30%	25%	100%	0.91(0.84-0.98)	98%	0.84(0.74-0.94)	90%	0.85(0.76-0.94)
	45%	100%	0.92(0.85-0.98)	98%	0.84(0.76-0.92	94%	0.84(0.76-0.92)
	65%	100%	0.92(0.85-0.98)	98%	0.92(0.88-0.96)	94%	0.88(0.80-0.95)
45%	25%	100%	0.92(0.56-0.98)	98%	0.97(0.91-1.0)	94%	0.99(0.97-1.0)
	45%	100%	0.86(0.72-0.94)	98%	0.97(0.93-0.99)	92%	0.98(0.94-0.99)
	65%	100%	0.91(0.85-0.98)	98%	0.99(0.96-0.99)	92%	0.93(0.86-0.99)
65%	25%	100%	0.90(0.83-0.97)	98%	0.96(0.91-0.99)	94%	0.88(0.81-0.98)
	45%	100%	0.90(0.81-0.98)	98%	0.97(0.93-0.99)	92%	0.89(0.81-0.97)
	65%	100%	0.91(0.84-0.98)	98%	0.95(0.89-0.99)	94%	0.82(0.73-0.92)

According to Table 4.9, the proposed framework correctly distinguished the number of clusters with a high degree of agreement for all types of missing values, and all rates of overlap. The lower agreement was observed under MNAR for 65% missingness and 65% overlap. Thus, it is concluded that our proposed framework could handle all kinds of missingness.

Third scenario; consensus clustering

According to the literature review, two methods of consensus clustering were presented, Majority Vote and Co-Membership (Section 3.3). As part of this study, 4M method was developed based on model-based clustering on multivariate multinomial distribution. Consequently, the objective of the third scenario was to compare three consensus clustering methods. So, once incomplete mixed datasets were generated, using different rates for overlapping and missingness, Step 1 to Step 4 of the proposed framework (Table 4.7) were applied. The results are the same at this point. Then, in Step 5, three consensus clustering methods Majority Vote, Co-Membership, and 4M were applied. The agreement results for these three consensus methods and the reference cluster were shown in Table 4.10.

It is surprising that none of the analyses performed with Majority Vote detected two clusters. After careful investigation, a single or two values in each imputed dataset were classified into the individual cluster. Therefore, due to Basagaña assumption, the number of clusters was not fixed for all imputed datasets, and the cluster analysis was not repeated with the fixed number of clusters, therefore, there are few datasets that distinguish two clusters in Majority Vote.

Consequently, since Kappa values were calculated for all of the imputed datasets that were classified on two or more than two clusters, it was not possible to report a high level of agreement. As a result, Majority Vote failed to identify the correct cluster number, and this result proved Majority Vote was not a useful consensus clustering method. In Co-Membership, there are situations in which the correct number of clusters could not be determined correctly, however, they are not as frequent as on Majority Vote.

Table 4.10: Kappa Cohen coefficients (95% CI) and percentage of the correct number of clusters derived for Third Scenario

Overlap (%)	Missing (%)	Majority Vote		Co-Membership		4M	
		Achieved Clusters(%)	Kappa (95 % CI)	Achieved Clusters(%)	Kappa (95 % CI)	Achieved Clusters(%)	Kappa (95 % CI)
30%	25%	0%	-	18%	0.75(0.68-0.81)	100%	0.91(0.84-0.98)
	45%	2%	0.30(0.27-0.34)	80%	0.75(0.70-0.79)	100%	0.92(0.85-0.98)
	65%	6%	0.31(0.27-0.34)	94%	0.68(0.66-0.75)	100%	0.92(0.85-0.98)
45%	25%	0%	-	11%	0.80(0.72-0.87)	100%	0.92(0.56-0.98)
	45%	0%	-	71%	0.61(0.56-0.68)	100%	0.76(0.72-0.84)
	65%	3%	0.29(0.25-0.33)	96%	0.61(0.59-0.64)	100%	0.91(0.85-0.98)
65%	25%	0%	-	24%	0.63(0.61-0.71)	100%	0.90(0.83-0.97)
	45%	0%	-	85%	0.54(0.51-0.56)	100%	0.90(0.81-0.98)
	65%	0%	-	96%	0.38(0.35-0.41)	100%	0.91(0.84-0.98)

When the rate of missing values increased Co-Membership method could determine the correct number of clusters more accurately. However, its agreement is not very high in this method. Although using 4M method, the number of clusters was determined exactly, in further investigation, the imputed datasets which were classified into three clusters included few objects, and the probability of these objects to create a separate mixture in 4M method was not significant and the final decision was made to count 2 clusters. Therefore, using 4M method, the objects were classified into two clusters that showed high agreement with the reference cluster.

Fourth scenario; packages comparison

The main purpose of the fourth scenario was to compare two complete clustering frameworks, namely the proposed framework (Table 4.7) and Basagaña framework (Table 4.2). Both frameworks started using multiple imputation and cluster analysis was applied by K-means, however, Basagaña framework selected the key important variables and the proposed framework is based on variable reduction. Therefore, two different mixed datasets with different numbers of variables were examined in this scenario. As Basagaña's package takes a long time, we only defined two values for overlapping rates and the rates for missingness rates. Table 4.11 shows the results of these comparisons.

Table 4.11: Kappa Cohen coefficients (95% CI) and percentage of the correct number of clusters derived for Fourth Scenario

Nr. Variables	Overlap (%)	Missing (%)	Basagaña framework		Proposed framework	
			Clusters (%)	Kappa (95 % CI)	Clusters (%)	Kappa (95 % CI)
56%	30%	10%	100%	0.88(0.79-0.96)	100%	0.89(0.81-0.96)
		20%	100%	0.86(0.76-0.96)	100%	0.89(0.81-0.96)
	65%	10%	100%	0.82(0.69-0.95)	100%	0.92(0.85-0.98)
		20%	100%	0.86(0.79-0.96)	100%	0.88(0.76-0.97)
86%	30%	10%	100%	0.87(0.77-0.96)	100%	0.88(0.80-0.95)
		20%	100%	0.86(0.77-0.95)	100%	0.87(0.79-0.94)
	65%	10%	100%	0.87(0.77-0.96)	100%	0.88(0.80-0.95)
		20%	100%	0.86(0.77-0.95)	100%	0.87(0.80-0.96)

For two considered numbers of variables, both methods recognized exactly the correct number of clusters. According to Kappa values and the corresponding 95% confidence interval, the proposed method assigns the clusters slightly better than Basagaña's method with no significant differences. The first point of this scenario is that the two introduced methods for dimension reduction have no strong effect on

the efficiency of the framework. In contrast to Table 4.10, Majority Vote consensus method performs properly as an integrated chain in Basagaña's package and this is a critical aspect of this scenario. However, the weakness of Basagaña method is that it is too time-consuming when the number of variables increases since this method is searching for the best combination of the number of clusters and variable selection and refits the cluster analysis on m imputed dataset. Although Basagaña's framework keeps the number of clusters fixed and selected the specific variables, while the proposed method reduces the number of variables and applies consensus clustering to an arbitrary number of clusters for each imputed dataset, both frameworks perform well with good classification results.

Fifth scenario; clustering methods

Among the various clustering methods, the combination of K-means and hierarchical clustering was utilized in the proposed framework, however, model-based clustering has been a frequently used and well-established method in the last decades. This scenario attempts to apply model-based clustering and BIC rather than the combination of hierarchical clustering and K-means to determine the number of clusters and identify the clusters. Therefore, in Step 3 of Table 4.7, the model-based clustering methods using different types of variance-covariance matrix structures were fitted on each imputed dataset to determine the best-fitted model and number of clusters using BIC. Then, in Step 4 of Table 4.7, the model-based clustering was refitted on each imputed dataset according to the best type of model and number of clusters to assign objects to clusters. The rest of the framework was the same as presented in Table 4.7. Table 4.12 provides the results of these two competing methods for clustering under various rates of overlapping and missingness.

In all cases, both the model-based clustering and the combination of K-means and hierarchical clustering correctly identified the number of clusters. However, the agreement between the reference clustering and results of the proposed framework was higher than the results observed with the model-based method. The simulated datasets were generated from mixture normal and multinomial distributions, so, the agreement between reference clustering and assigned clustering in the model-based method was expected to be higher and closer to the results of the proposed method. However, based on the results in Table 4.12, it appears that after variable reduction in Step 2 of Table 4.7, the distribution of principal components was changed, and the changes in the original variables had a significant effect on the models fitted to the

Table 4.12: Kappa Cohen coefficients (95% CI) and percentage of the correct number of clusters derived for Fifth Scenario

		Model-Based clustering		Proposed Method	
Overlap (%)	Missing (%)	Clusters (%)	Kappa (95 % CI)	Clusters (%)	Kappa (95 % CI)
30%	25%	100%	0.69(0.628-0.75)	100%	0.91(0.84-0.98)
	45%	100%	0.65(0.61-0.69)	100%	0.92(0.85-0.98)
	65%	100%	0.61(0.59-0.68)	100%	0.92(0.85-0.98)
45%	25%	100%	0.71(0.63-0.78)	100%	0.92(0.56-0.98)
	45%	100%	0.58(0.53-0.65)	100%	0.76(0.72-0.84)
	65%	100%	0.55(0.53-0.58)	100%	0.91(0.85-0.98)
65%	25%	100%	0.60(0.58-0.68)	100%	0.90(0.83-0.97)
	45%	100%	0.50(0.47-0.53)	100%	0.90(0.81-0.98)
	65%	100%	0.66(0.57-0.75)	100%	0.91(0.84-0.98)

new components in imputed datasets. Therefore, as a result of the variable reduction, objects are assigned to clusters in model-based clustering with the lower agreement. Consequently, according to the simulated dataset, K-means and hierarchical clustering achieved better results than model-based clustering.

4.4.4 Conclusion

There are many methods to handle missing values, variable reduction, and cluster analysis separately. In chapter 2 and chapter 3, several of these methods required for the current full frameworks and for the proposed framework were explained. In this chapter, these methods were integrated using different combinations. In these extensive simulations, by designing different scenarios in handling missing values, dimension reduction, and employing several methods and packages for clustering, we attempted to show the influence of the different methods on the result of the correct number of clusters and also assigning the objects into clusters, comprehensively.

Along with comprehensive comparisons, the effectiveness of the methods was evaluated. Based on the simulation study, it was necessary to impute missing values when applying K-means and hierarchical methods. In addition, the proposed method could perform efficiently on all types of missingness. If model-based clustering is replaced by K-means and hierarchical methods in the proposed framework, it would have a slightly lower ability in classification. The proposed method for consensus clustering is highly efficient as compared to existing approaches. Basagaña's approach works usefully just in its own package by using a fixed number of clusters

and selected key variables. However, the package is too time-consuming and computation time increases as the number of objects and variables increases.

In conclusion, the proposed framework is efficiently designed on cluster analysis using multiple imputation in multidimensional data and has an effective performance compared to competing methods based on the designed scenarios on these simulated datasets.

CHAPTER 5

Clustering on COPD patients

In Chapter 1, COPD was introduced as a disease with a large number of phenotypes and a complex and heterogeneous multifactorial background. The goal of this chapter is to identify clinical phenotypes in adults suffering from COPD through cluster analysis. Using a combination of the methods introduced in the previous two chapters, this section performed cluster analysis on the multidimensional COPD dataset which deals with missing values using multiple imputation. Detailed descriptions of the COPD dataset and the variables, as well as the percent of missing values, are presented in Section 5.1. A clustering framework for the multidimensional COPD dataset, using multiple imputation and clustering methods, is proposed in Section 5.2. The results of clustering are presented in Section 5.3. Section 5.4 contains a comprehensive discussion on the results of clustering and clinical interpretation.

This chapter is based on

Nekoee Zahraei, H., Guissard, F., Paulus, V., Henket, M., Donneau, A.F., & Louis, R. (2020). Comprehensive Cluster Analysis for COPD Including Systemic and Airway Inflammatory Markers. COPD: Journal of Chronic Obstructive Pulmonary Disease, 1 – 12. https://doi.org/10.1080/15412555.2020.1833853)

5.1 COPD dataset

This chapter examined clustering among 178 COPD patients in a stable state recruited from ambulatory care. All patients were above 40 years, had a smoking history of more than 20 pack-years, post-bronchodilator FEV1/FVC < 70%, and denied any history of asthma before 40 years.

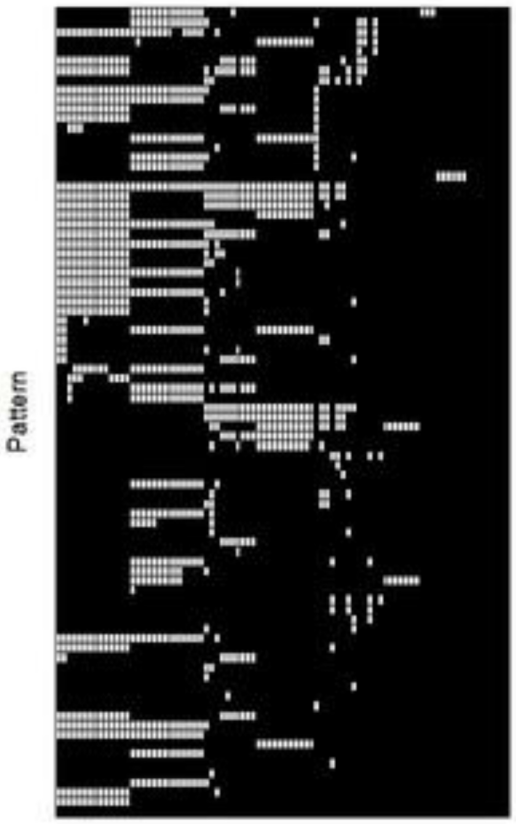

Figure 5.1: A general overview of COPD data; white cells are missing values (n = 178 patients)

In this study, the patients were described by a total of 84 mixed sets of variables with missing values. Demographic variables, pulmonary function tests, treatment features, blood cell counts and systemic inflammatory markers, atopic status, sputum cell counts, and microbiology were included in the study and discussed in detail in this section. In usual studies, the clusters were based on demographic variables, symptoms, spirometry, imaging, and comorbidities. Clustering has not been widely used in studies investigating the airway inflammatory component and the atopic status in large cohort studies of COPD. In our standard routine investigation of COPD, we included serum IgE, blood eosinophils and sputum eosinophils as well as FeNO as markers to assess the T2 trait and sought to see whether T2 trait is common and

strong enough to shape a cluster in a population of COPD that denies any history of asthma before 40 years of age.

Characteristics of the patients and the percentage of missing values before imputing are presented in Table 5.1. Quantitative variables were summarized using median and interquartile range (P25 - P75); while count and percentage were used for qualitative variables. In addition, in Figure 5.1, a general perspective of COPD dataset was presented. There are columns that represent variables and rows that represent COPD patients. The white cells are representative of missing data.

More than half of the patients were males (54.49%) with age ranging from 40 to 84 years and the median age is 64 years. Patients displayed a normal weight (median body mass index was $23.62 kg/m^2$). They had a consistent tobacco consumption history with a median pack/year of 37 during median 43 years smoking duration.

A total of 61% of patients have not received an OCS course the year prior to the visit, while 24% took it once, and 15% took it more than once. A third of patients have not received antibiotics while 85% took antibiotics once; about the same percentage have never been admitted to the emergency room for asthma or COPD and also have never been hospitalized for asthma or COPD, too. The type of Missing values were missing completely at randoms and the percentage of missing values ranged from 0% to 23% and 75% of patients presented at least one missing value.

5.2 Clustering Framework

Outlines of the applied framework to assign clustering for COPD patients are described in Table 5.2. Missing values were imputed by drawing from the posterior predictive distribution of Bayesian model. Multiple imputation replaces each missing value with a set of 100 plausible values. Then, 100 imputed datasets are analyzed by the method that would have been appropriate if the data had been complete. Since this study contains quantitative and qualitative variables, FAMD (Section 3.2.2) was applied in each imputed dataset for creating new components. Then, the percentage of contribution of variables was determined.

Table 5.1: Characteristics of the COPD dataset ($n = 178$)

Variable	**Median (IQR)** **Percentage (frequency)**	**missing value** **%(n)**
Demographic characteristics		
Age (Year)	64.5(57 - 72)	0% (0)
Sex (Male)	54.5% (97)	0% (0)
Height (cm)	167(160 - 175)	0.56% (1)
Weight (kg)	67(58 - 78)	0.56% (1)
BMI (kg/m^2)	23.6 (21.22 - 27.18)	0.56% (1)
Cigarette Packs (Year)	37.2 (22.5 - 50)	2.25% (4)
Cigarettes (day)	20 (10 - 23.5)	3.93% (7)
Smoking Duration(Year)	43 (33 - 50)	2.8% (5)
OCS Course		5.62% (10)
0	60.7% (102)	
1	23.8% (40)	
≥2	15.5% (26)	
Antibiotic Course		4.49% (8)
0	36.5% (62)	
1	51.8% (88)	
≥2	11.8% (20)	
Emergency Room Admission for asthma or COPD		1.12% (2)
0	84.7% (149)	
1	14.8% (26)	
≥2	0.6% (1)	
Number of hospitalizations for asthma or COPD		2.8% (5)
0	85.5% (148)	
1	12.7% (22)	
≥2	1.7% (3)	
Pulmonary characteristics		
FeNO (ppb)	16 (10 - 25)	7.30% (13)
FEV1 (mL)	1380 (1085 - 1810)	0% (0)
FEV1 (% predicted)	53 (43 - 66)	0% (0)
FEV1 PD (mL)	1480 (1180 - 1930)	0.56% (1)
FEV1 PD (% predicted)	57 (47 - 71)	0.56% (1)
Reversibility (%)	7 (1 - 12)	0.56% (1)
FVC (mL)	2505 (2020 - 3067.5)	0% (0)
FVC (% predicted)	77 (64 - 89)	0% (0)
FVC post (mL)	2580 (2080 - 3200)	0.56% (1)
FVC post (%)	80 (68 - 92)	0.56% (1)
FEV1/ FVC pre (%)	56.4 (48.1 - 64.3)	0% (0)
FEV1/ FVC post (%)	57.9 (49.4 - 66.5)	0.56% (1)
TLC (mL)	6.5 (5.4 – 7.2)	16.29% (29)
TLC (%predicted)	109 (96 - 123)	16.29% (29)
RV (mL)	3.6 (3.1 – 4.5)	16.29% (29)
RV (% predicted)	168 (137 - 201)	16.29% (29)
RV/TLC (%)	59.6 (52.8 – 66.2)	16.29% (29)
DLCO (mmol/kPa.min)	3.9 (2.9 - 5.3)	17.41% (31)
DLCO (% predicted)	49 (37.5 - 60)	17.41% (31)
DLCO/VA(mmol/kPa.min/l)	0.9 (0.7 - 1.2)	17.41% (31)
DLCO/VA (% predicted)	67 (50 – 81.5)	17.41% (31)
sGaw (1/kPa*sec)	0.5 (0.4 - 0.7)	23.03% (41)
FRC (L)	4.9 (4.1 - 5.9)	21.34% (38)
FRC (% predicted)	158 (138 - 183)	21.34% (38)
Treatment characteristics		
Treatment (Yes)	61.8% (105)	4.49% (8)
ICS (Yes)	55.1% (97)	1.12% (2)
OCS (Yes)	5.1% (9)	1.12% (2)
LAMA (Yes)	51.1% (90)	1.12% (2)
LABA (Yes)	67% (118)	1.12% (2)
SABA (Yes)	41% (72)	1.12% (2)
LTRA (Yes)	3.4% (6)	1.12% (2)
Theophylline (Yes)	2.8% (5)	1.12% (2)
CAT score	25 (16 – 31)	0% (0)

Table 5.1 – continued from previous page

Variable	Median (IQR) Percentage (frequency)	missing value %(n)
Blood characteristics		
Leucocytes (1/μl)	7.9 (6.7 - 9.7)	7.86% (14)
Neutrophils (%)	60.7 (54 - 67.7)	8.43% (15)
Lymphocytes (%)	27.8 (21.3 - 34.3)	8.43% (15)
Monocytes (%)	7.9 (6.5 - 9.4)	8.43% (15)
Eosinophils (%)	2 (1 - 3.2)	8.43% (15)
Basophils (%)	0.4 (0.3 - 0.6)	8.43% (15)
Neutrophils (1/μl)	4840.7 (3782.5 - 6027.3)	8.43% (15)
Lymphocytes (1/μl)	2213.4 (1739.1 - 2710.8)	8.43% (15)
Monocytes (1/μl)	646.9 (496.1 - 836.3)	8.43% (15)
Eosinophils (1/μl)	147.6 (77.5 - 252.0)	8.43% (15)
Basophils (1/μl)	31.2 (21.4 - 48.3)	8.43% (15)
Fibrinogen (g/l)	3.5 (3 - 4)	4.49% (8)
CRP (mg/l)	2.5 (1.1 - 5.7)	4.49% (8)
Alpha 1 antitrypsin (g/l)	1.5 (1.4 - 1.7)	12.36% (22)
Calcium (mmol/L)	2.4 (2.4 - 2.5)	6.74% (12)
25(OH) Vitamine D (ng/ml)	20 (12 - 31)	13.48% (24)
Phosphate (mmol/L)	0.9 (0.8 - 1.1)	6.74% (12)
Atopy characteristics		
Total IgE (KU/L)	72.5 (22.7 - 227.7)	10.11% (18)
RAST DPT (d1) %>0.35 (KU/L)	11.4% (18)	11.24% (20)
RAST Cat (e1), %>0.35 (KU/L)	3.7% (6)	10.11% (18)
RAST Dog (e5), %>0.35 (KU/L)	3.1% (5)	10.11% (18)
RAST Grass (GX3), %>0.35 (KU/L)	7% (11)	11.24% (20)
RAST microog (MIX1), %>0.35 (KU/L)	9.4% (15)	10.11% (18)
RAST Birch (t3), %>0.35 (KU/L)	1.2% (2)	10.67% (19)
Sputum characteristics		
Positive Aerobic Sputum Culture	9% (13)	18.54% (33)
Weight of sputum (g)	1.7 (1.1 - 2.9)	20.78% (37)
Total Cell Counts (10^6/g)	2.3 (0.9 - 5.6)	20.78% (37)
Squamous (%)	10 (3 - 33)	20.78% (37)
Viability (%)	69 (55 - 84)	20.78% (37)
Macrophages (%)	12.3 (5 - 23.6)	21.91% (39)
Lymphocytes (%)	1 (0 - 2)	21.91% (39)
Neutrophils (%)	74.8 (56.87 - 91.05)	21.91% (39)
Eosinophils (%)	1.3 (0.2 - 4.4)	21.34% (38)
Epithelial cells (%)	2.5 (0.6 - 7.7)	21.91% (39)
Macrophages (10^3/g)	268.3 (80 - 583.8)	23.03% (41)
Lymphocytes (10^3/g)	18.5 (0 - 43.8)	23.03% (41)
Neutrophils (10^3/g)	1307.5 (503.9 - 3831.6)	22.47% (40)
Eosinophils (10^3/g)	28.9 (2.8 - 232.7)	22.47% (40)
Epithelial cells (10^3/g)	55.4 (11.2 - 202.9)	22.47% (40)

[1] BMI(Body Mass Index); OCS(Oral Corticosteroids); FENO(Fractional Exhaled Nitric Oxide); FEV1(Forced Expiratory Volume in one second); FVC(Forced Vital Capacity); TLC(Total Lung Capacity); RV(Residual Volume); DLCO(Diffusing Capacity for Carbon Monoxide); FRC(Functional residual capacity); LABA(Long Acting B2 Agonist); LAMA(Long Acting Muscarinic Antagonist); LTRA(Leukotriene Receptor Antagonist); CRP(C-Reactive Protein); CAT(COPD Assessment Test), ICS(Inhaled Corticosteroids)

In the cluster analysis step, the number of clusters for each imputed dataset was determined by hierarchical clustering and a package of 30 indices for determining the relevant number of clusters, then K-means was applied for assigning clusters to COPD patients. The derived results from the 100 analyses are then combined to produce the final quantity of interest following Rubinâs rules. In the consensus step, the final clustering result was achieved using the co-membership (Section 3.3.2) method by minimizing the sum of the squared distance of existing clustering results.

In this step, for each clustering output, two indices for internal clustering validation and stability validation were calculated. The output of consensus clustering was considered as the individual final clustering result for the original incomplete COPD

dataset and all the imputed datasets. Then, median and interquartile ranges were calculated for all variables in the original incomplete COPD dataset and for each imputed dataset. Finally, an overall median with the corresponding interquartile range was calculated over all 100 imputed datasets.

All variables were compared between the derived clusters using Kruskal-Wallis and Chi-squared tests for quantitative and qualitative variables, respectively. Comparison among two clusters was applied according to Dunnâs multiple comparison test. There are no significant differences between two clusters with the same letter. Finally, the difference between the three clusters was displayed by boxplots. All analyses were performed using R statistical software. P-values < 0.05 were considered statistically significant.

Table 5.2: Framework of variable reduction and cluster analysis to determine clusters in COPD

Input: COPD dataset with missing values and multidimensional variables
Step 1. Multiple Imputation
i) Obtain 100 complete datasets by multiple imputation (MICE)
Step 2. Factor analysis for mixed data (FAMD)
i) Determine quantitative and qualitative variable
ii) Apply FAMD for each imputed dataset
iii) Determine the number of components for each imputed dataset
Step 3. Hierarchical clustering
i) Choosing the best number of clusters for each imputed dataset
Step 4. Partitioning Clustering
i) Consider the number of clusters determined in the previous step
ii) Assign patients to each cluster in each imputed dataset by partitioning clustering
Step 5. Co-Membership method for Consensus Clustering
i) Combine all ensemble clustering to get a final best clustering
Output: Partition of clustering labels $C = \{C_1, \ldots, C_k\}$
Step 6. Assign patients to the final result of consensus clustering
i) Allocate patients in the original incomplete dataset and each imputed dataset to calculate the final result of consensus clustering
Step 7. Description of clustering
i) Calculate median for the original dataset and overall median for imputed datasets
ii) Comparison between cluster (Kruskal-Wallis and Chi-squared tests and Dunn's multiple comparison test)
Output: Descriptive analysis tables for clustering

MICE, *FactoExtra*, *FactoMineR*, and clue are R packages that implement the mul-

tiple imputation method, FAMD, cluster analysis, and Co-Membership method for consensus clustering. This is based on all operational processes listed in Table 5.2.

5.3 Clustering Results

Following the framework explained in the previous section, the following results were obtained., After variable reduction using FAMD, was possible to calculate the percentage contribution of variables for the next clustering step. Table 5.3 shows the order and the impact of each variable on clustering. The highest contribution for variables in clustering was for FEV1 (mL), FEV1 PD (mL), FEV1 PD (% predicted), FEV1 (% predicted), and FVC (mL). The order of the contributions of variables from highest to lowest is shown in Figure 5.2.

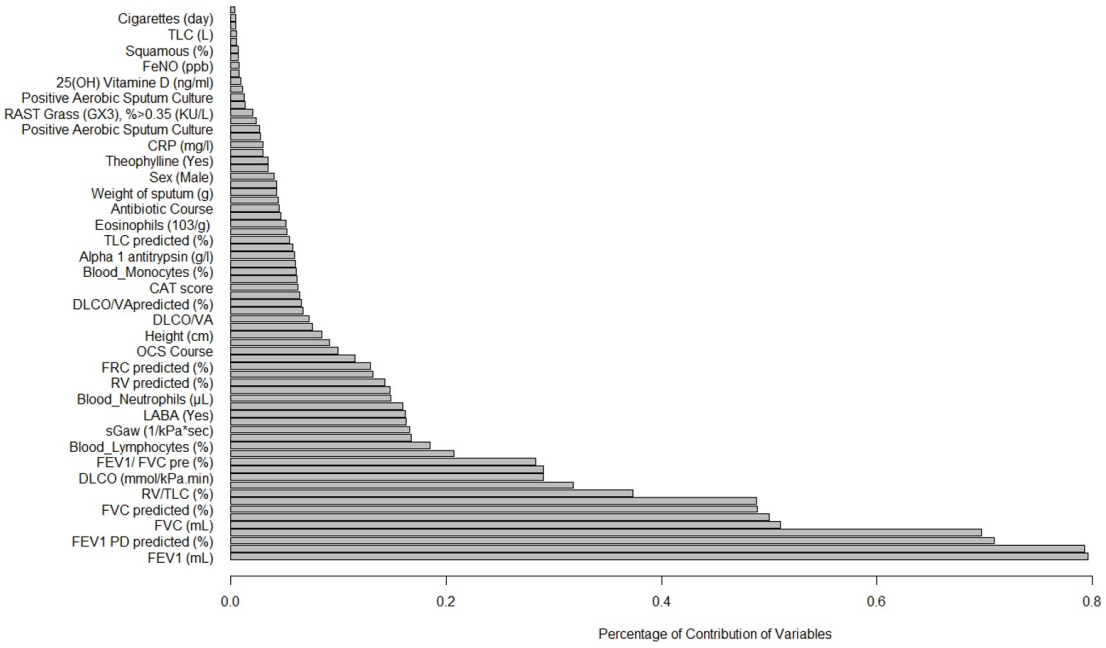

Figure 5.2: The percentage contribution of COPD variables in principal components from the highest to lowest in the whole cohort (n = 178 patients)

After applying multiple imputation, 83 imputed datasets suggested 35 new components, and the others, 17%, were summarized into 36 components. Next, each imputed dataset was classified based on its own selected components. A total of 97 out of 100 imputed datasets were classified into 3 clusters, and 3% into 4 clusters.

Table 5.3: The percentage contribution of COPD variables in principal components

Order	Variables	Percentage of contribution value	Order	Variables	Percentage of contribution value
1	FEV1 (mL)	0.80	43	Sputum_Eosinophils (103/g)	0.05
2	FEV1 PD (mL)	0.79	44	SABA (Yes)	0.05
3	FEV1 PD (% predicted)	0.71	45	Antibiotic Course	0.05
4	FEV1 (% predicted)	0.70	46	Age (Year)	0.04
5	FVC (mL)	0.51	47	Weight of sputum (g)	0.04
6	FVC post (mL)	0.50	48	Sputum_Neutrophils (103/g)	0.04
7	FVC (% predicted)	0.49	49	Sex (Male)	0.04
8	FVC post (%)	0.49	50	Sputum_Lymphocytes (103/g)	0.04
9	RV/TLC (%)	0.37	51	Theophylline (Yes)	0.03
10	FEV1/ FVC post (%)	0.32	52	Smoking Duration(Year)	0.03
11	DLCO (mmol/kPa.min)	0.29	53	CRP (mg/l)	0.03
12	Emergency Room	0.29	54	FRC (L)	0.03
13	FEV1/ FVC pre (%)	028	55	Positive Aerobic Sputum Culture	0.03
14	Blood_Neutrophils (%)	0.21	56	Blood_Eosinophils (%)	0.02
15	Blood_Lymphocytes (%)	0.19	57	RAST Grass (GX3), %>0.35 (KU/L)	0.02
16	DLCO (% predicted)	0.17	58	Sputum_Neutrophils (%)	0.01
17	sGaw (1/kPa*sec)	0.17	59	Positive Aerobic Sputum Culture	0.01
18	LAMA (Yes)	0.16	60	IgE (KU/L)	0.01
19	LABA (Yes)	0.16	61	25(OH) Vitamine D (ng/ml)	0.01
20	ICS (Yes)	0.16	62	Blood_Basophils (%)	0.01
21	Blood_Neutrophils (1/μl)	0.15	63	FeNO (ppb)	0.01
22	Treatment (Yes)	0.15	64	RAST Birch (t3), %>0.35 (KU/L)	0.01
23	RV (% predicted)	0.14	65	Squamous (%)	0.01
24	Number of hospitalizations	0.13	66	Sputum_Epithelial cells (%)	0.01
25	FRC (% predicted)	0.13	67	TLC (mL)	0.01
26	Weight (kg)	0.12	68	Sputum_Eosinophils (%)	0.01
27	OCS Course	0.10	69	Cigarettes (day)	0.01
28	RV (mL)	0.09	70	Blood_Monocytes (1/μl)	0.004
29	Height (cm)	0.09	71	RAST DPT (d1) %>0.35 (KU/L)	0.003
30	Blood_Lymphocytes (1/μl)	0.08	72	Sputum_Lymphocytes (%)	0.003
31	DLCO/VA(mmol/kPa.min/l)	0.07	73	RAST moul (MIX1), %>0.35	0.002
32	BMI (kg/m2)	0.07	74	Sputum_Macrophages (103/g)	0.002
33	DLCO/VA (% predicted)	0.07	75	Phosphate (mmol/L)	0.001
34	Blood_Leucocytes (1/μl)	0.06	76	RAST Cat (e1), %>0.35 (KU/L)	0.001
35	CAT score	0.06	77	LTRA (Yes)	0.001
36	OCS (Yes)	0.06	78	Sputum_Macrophages (%)	0.001
37	Blood_Monocytes (%)	0.06	79	Blood_Basophils (1/μl)	0.001
38	Total Cell Counts (106/g)	0.06	80	RAST Dog (e5), %>0.35 (KU/L)	0.001
39	Alpha 1 antitrypsin (g/l)	0.06	81	Calcium (mmol/L)	0.001
40	Fibrinogen (g/l)	0.06	82	Blood_Eosinophils (1/μl)	0.001
41	TLC (%predicted)	0.05	83	Reversibility (%)	0.001
42	Viability (%)	0.05	84	Cigarette Packs (Year)	0.001

Finally, after completing the framework presented in Table 5.2 using consensus clustering, three distinct clusters with acceptable values for validation were identified. According to Table 5.4, two indices for internal clustering validation and two indices for clustering stability validation are reported. Silhouette coefficient and Dunn index values for internal measures are 0.61 and 0.54, respectively. The average proportion of non-overlap and the average distance between means for stability measures are 0.02 and 0.01, respectively.

Table 5.4: Clustering validation for the COPD dataset (n=178)

	Indices	Value
Internal measures	Silhouette Coefficient	0.61
	Dunn Index	0.54
Stability measures	Average Proportion of Non-overlap	0.021
	Average Distance between Means	0.008

The clustering results for the original incomplete COPD dataset and overall clustering result for 100 imputed datasets are displayed in Table 5.5 and Table 5.6, respectively. Three different clusters, which shared similar smoking history were found. Cluster 1 included men with moderate airway obstruction ($n = 67$) while cluster 2 comprised men who were exacerbation-prone, with severe airflow limitation and intense granulocytic airway and neutrophilic systemic inflammation ($n = 56$). Cluster 3 essentially included women with moderate airway obstruction ($n = 55$). All clusters had a low rate of bacterial colonization ($< 5\%$), a low median FeNO value ($< 20 ppb$), and a very low sensitization rate towards common aeroallergens ($0-15\%$). CAT score did not differ between clusters.

There were striking sex differences between the clusters with a clear dominance of males in clusters 1 and 2 while cluster 3 was essentially composed of women. Smoking history was similar between clusters and BMI was slightly higher in cluster 1 while still remaining in the normal range. Clusters 2 and 3 received more frequent courses of OCS the year prior to the visit whereas there was no difference regarding the number of the antibiotic courses. Clusters 2 and 3 were those in which patients received more maintenance treatment including ICS, LAMA, LABA, and also used more often SABA as a reliever. Interestingly, CAT score did not differ between the three clusters despite clear differences in lung function impairment (Figure 5.3).

Table 5.5: Characteristics of patients with COPD before imputation, Median (IQR) / Percentage (frequency) in each cluster, and comparison between clusters

Result for clustering on the original COPD dataset				
Variable	**Cluster 1 (n=67)**	**Cluster 2 (n=56)**	**Cluster 3 (n=55)**	**P-value**
Demographic characteristics				
Age (Year)	62(55-66)^{b}	67(60.7-74.2)^{a}	67(58.5-73)^{a}	<0.001
Sex (Male)	76.1%(51[67])^{a}	75%(42[56])^{a}	7.3%(4[55])^{b}	<0.0001
Height (cm)	174(167-178)^{a}	169.5(162-175.2)^{a}	161(156-165)^{b}	<0.0001
Weight (kg)	77(65-89)^{c}	67(56.7-76)^{a}	60(53-66)^{b}	<0.0001
BMI (kg/m^2)	25.6(22.2-30.2)^{b}	23.1(20.6-24.9)^{a}	23.1(19.9-25.5)^{a}	<0.001
Cigarette Packs (Year)	36.9(21-50)^{a}	42.5(23.2-52.5)^{a}	35(22-43.9)^{a}	>0.05
Cigarettes (day)	20(10-25)^{a}	20(10-2.25)^{a}	20(10-20)^{a}	>0.05
Smoking Duration(Year)	40(32-46.5)^{c}	47(34.2-52.2)^{a}	44(37-50)^{ab}	<0.05
OCS Course				<0.0001
0	81.2%(52[64])	31.4%(16[51])	64.1%(34[53])	
1	14.1%(9[64])	35.3%(18[51])	24.5%(13[53])	
≥ 2	4.7%(3[64])	33.3%(17[51])	11.3%(6[53])	
Antibiotic Course				>0.05
0	42.9%(27[63])	32%(17[53])	33.3%(18[54])	
1	47.6%(30[63])	50.9%(27[53])	57.4%(31[54])	
≥ 2	9.5%(6[63])	16.9%(9[53])	9.3%(5[54])	
Emergency Room Admission for asthma or COPD				<0.0001
0	95.4%(63[66])	70.9%(39[55])	85.4%(47[55])	
1	4.5%(3[66])	27.3%(15[55])	14.5%(8[55])	
≥ 2	0%(0[66])	1.8%(1[55])	0%(0[55])	
Number of hospitalizations for asthma or COPD				<0.0001
0	97%(64[66])	67.9%(36[53])	88.9%(48[54])	
1	3%(2[66])	26.4%(14[53])	11.1%(6[54])	
≥ 2	0%(0[66])	1.9%(1[53])	0%(0[54])	
Pulmonary characteristics				
FeNO (ppb)	15(10.5-23)^{a}	19(10-32.25)^{a}	15(8-24)^{a}	>0.05
FEV (mL)	1950(1700-2205)^{b}	1085(780-1352.5)^{a}	1240(1075-1390)^{a}	<0.0001
FEV1 (% predicted)	66(56-78)^{c}	37(30-45.2)^{a}	55(47.5-66)^{b}	<0.0001
FEV1 PD (mL)	2000(1820-2405)^{b}	1190(837.5-1380)^{a}	1310(1135-1470)^{a}	<0.0001
FEV1 PD (% predicted)	69(60.5-81.5)^{c}	40(32.7-50)^{a}	59(51.5-69.5)^{b}	<0.0001
Reversibility (%)	6(0.5-10.5)^{a}	6.5(1-11.25)^{a}	8(2-13.5)^{a}	>0.05
FVC (mL)	3200(2625-3780)^{c}	2427.5(1705-2767.5)^{a}	2140(1820-2390)^{b}	<0.0001
FVC (% predicted)	86(74-99.5)^{c}	65(55.7-74.2)^{a}	77(68.5-85.5)^{b}	<0.0001
FVC post (mL)	3320(2805-3870)^{c}	2385(1797.5-2900)^{a}	2130(1870-2445)^{b}	<0.0001
FVC post (%)	90(79.5-101)^{c}	68(56.7-76.5)^{a}	80(71.5-91.5)^{b}	<0.0001
FEV1/ FVC pre (%)	63.1(56.5-66.6)^{b}	47(40.4-50.4)^{a}	57.1(53.1-64.5)^{b}	<0.0001
FEV1/ FVC post (%)	63.9(58.3-68.5)^{b}	47.5(42.8-53.2)^{a}	59.9(54.5-67)^{b}	<0.0001
TLC (mL)	6710(5725-7385)^{a}	6945(5832.5-8107.5)^{a}	5430(4890-6005)^{b}	<0.0001
TLC (% predicted)	103(93-112.5)^{b}	112.5(100-125.2)^{a}	117(102-127)^{a}	<0.0001
RV (mL)	3450(2990-3820)^{b}	4640(3687.5-5277.5)^{a}	3250(2845-3690)^{b}	<0.0001
RV (% predicted)	141(128-173)^{c}	182.5(162.5-228.5)^{a}	168(145.5-189.5)^{b}	<0.0001
RV/TLC (%)	53.2(47.1-56.8)^{c}	66.2(61.3-69)^{a}	60.95(56.9-64)^{b}	<0.0001
DLCO (mmol/kPa.min)	5(3.8-5.9)^{b}	3.5(2.7-4.5)^{a}	3.27(2.67-4)^{a}	<0.0001
DLCO (% predicted)	56(43-67)^{b}	42.5(30.7-57.2)^{a}	45(38-55)^{a}	<0.0001
DLCO/AV (mmol/kPa.min/l)	1(0.8-1.3)^{b}	0.8(0.6-1)^{a}	0.93(0.7-1.1)^{a}	<0.0001
DLCO/AV (% predicted)	70(56.5-88)^{b}	66(45.7-79)^{a}	62(50-75)^{a}	<0.001
sGaw (1/kPa*sec)	0.68(0.47-0.93)^{c}	0.36(0.24-0.47)^{a}	0.47(0.3-0.8)^{b}	<0.0001
FRC (L)	4.9(4.09-5.44)^{c}	5.9(4.9-6.5)^{a}	4.24(3.8-5.05)^{b}	<0.0001
CAT score	22(15-31)^{a}	26(19-30)^{a}	24(17-30)^{a}	>0.05
Treatment characteristics				
Treatment (Yes)	43.75%(28[64])^{a}	78.84%(41[52])^{b}	66.67%(36[54])^{b}	<0.0001
ICS (Yes)	30.30%(20[66])^{a}	70.91%(39[55])^{b}	69.09%(38[55])^{b}	<0.0001
OCS (Yes)	1.51%(1[66])^{a}	14.54%(8[55])^{b}	0%(0[55])^{a}	<0.0001
LAMA (Yes)	34.85%(23[66])^{a}	72.73%(40[55])^{b}	49.09%(27[55])^{a}	<0.0001
LABA (Yes)	45.45%(30[66])^{a}	83.64%(46[55])^{b}	76.36%(42[55])^{a}	<0.0001
SABA (Yes)	25.76%(17[66])^{a}	50.91%(28[55])^{b}	49.09%(27[55])^{b}	<0.0001
LTRA (Yes)	3.03%(2[66])^{a}	0%(0[55])^{a}	7.27%(4[55])^{a}	>0.05
Theophylline (Yes)	1.51%(1[66])^{a}	5.45%(3[55])^{a}	1.82%(1[55])^{a}	>0.05

Table 5.5 – continued from previous page

Result for clustering on the original COPD dataset				
Variable	**Cluster 1 (n=67)**	**Cluster 2 (n=56)**	**Cluster 3 (n=55)**	**P-value**
Blood characteristics				
Leucocytes (1/μl)	7.44(6.44-9.43)^a	8.58(7.02-10.75)^a	7.77(6.64-9.2)^a	>0.05
Neutrophils (%)	58.9(52.75-63.75)^b	68.6(61.27-73.22)^a	57.7(53.25-64.55)^b	<0.0001
Lymphocytes (%)	30.1(25.65-35)^b	20.7(15.7-26.85)^a	29(23.85-35.15)^b	<0.0001
Monocytes (%)	8.7(7-9.8)^b	7.8(6.37-9.82)^a	7.1(6.4-8.75)^a	<0.05
Eosinophils (%)	2.1(1.1-3.25)^a	1.8(0.85-3.12)^a	2.1(0.95-3.35)^a	>0.05
Basophils (%)	0.4(0.3-0.6)^a	0.3(0.2-0.52)^a	0.4(0.3-0.7)^a	>0.05
Neutrophils (1/μl)	4503.6(3554.3-5501.7)^b	5574.9(4116.2-7842.7)^a	4840.71(3393.7-5743.1)^b	<0.05
Lymphocytes (1/μl)	2346.79(1853.9-3019.6)^b	1885.8(1294.4-2424.2)^a	2267.98(1846.6-2710.8)^b	<0.0001
Monocytes (1/μl)	668.82(494.15-857.91)^a	707.29(534.38-864.82)^a	605.82(457.16-761.76)^a	>0.05
Eosinophils (1/μl)	150.72(84.57-240.24)^a	169.57(75.89-275.16)^a	147.63(85.47-248.25)^a	>0.05
Basophils (1/μl)	30.84(21.37-42.16)^a	27.63(18.64-47.66)^a	32.04(23.95-49.36)^a	>0.05
Fibrinogen (g/l)	3.42(2.87-3.86)^b	3.71(3.08-4.58)^a	3.54(2.91-3.80)^a	<0.05
CRP (mg/l)	2.3(1.25-5.6)^a	2.9(0.9-7.7)^a	2.1(1-5.15)^a	>0.05
Alpha 1 antitrypsin (g/l)	1.49(1.35-1.57)^b	1.63(1.39-1.76)^a	1.45(1.35-1.65)^b	<0.05
Calcium (mmol/L)	2.42(2.37-2.47)^a	2.41(2.35-2.48)^a	2.44(2.38-2.48)^a	>0.05
25(OH) Vitamine D (ng/ml)	19(12-32)^a	16.5(10-24.25)^a	24(13-32.5)^a	>0.05
Phosphate (mmol/L)	0.89(0.79-0.95)^a	0.87(0.71-1.03)^a	1.01(0.93-1.12)^b	<0.0001
Atopy characteristics				
IgE (KU/L)	52(23-233.5)^a	91.5(25-300.25)^a	73(19-170)^a	>0.05
RAST DPT (d1) %>0.35 (KU/L)	13.33%(8)^a	14.28%(7)^a	6.12%(3)^a	>0.05
RAST Cat (e1), %>0.35 (KU/L)	1.64%(1)^a	8.16%(4)^a	2%(1)^a	>0.05
RAST Dog (e5), %>0.35 (KU/L)	3.28%(2)^a	6.12%(3)^a	0%(0)^a	>0.05
RAST Grass (GX3), %>0.35 (KU/L)	6.67%(4)^a	12.24%(6)^a	2.04%(1)^a	>0.05
RAST microog (MIX1), %>0.35 (KU/L)	6.56%(4)^a	12.24%(6)^a	10%(5)^a	>0.05
RAST Birch (t3), %>0.35 (KU/L)	0%(0)^a	4.08%(2)^a	0%(0)^a	>0.05
Sputum characteristics				
Positive Aerobic Sputum Culture	5.26%(3)^a	13.95%(6)^a	8.89%(4)^a	>0.05
Weight of sputum (g)	1.95(1.37-3.88)^a	1.46(1.06-3.07)^a	1.64(0.93-3.94)^a	>0.05
Total Cell Counts (10^6/g)	1.46(0.96-2.81)^b	4.15(1.50-15.96)^a	1.75(0.98-3.43)^b	<0.05
Squamous (%)	21(6.5-42)^b	7.5(2-38.25)^a	7(2-14.5)^a	<0.05
Viability (%)	67(50-81)^a	69.5(51-86.25)^a	61(41-77.5)^a	>0.05
Macrophages (%)	14(5.8-28.15)^b	6.25(2.2-15.45)^a	18(9.25-35.65)^b	<0.0001
Lymphocytes (%)	1(0-2.1)^a	0.6(0.15-1.57)^a	1(0-2.6)^a	>0.05
Neutrophils (%)	70.6(53.3-88.2)^a	77.5(55.4-92.25)^a	52(37.6-80)^b	<0.001
Eosinophils (%)	1.6(0.2-5.45)^a	2.3(0.35-10.95)^a	2.8(0.9-20.6)^a	>0.05
Epithelial cells (%)	1.8(0.4-7.1)^a	2.6(0.57-10.97)^a	2.4(1.2-8.35)^a	>0.05
Macrophages (10^3/g)	231.57(40.01-564.3)^a	317.39(32.25-801.83)^a	154(15.65-557.81)^a	>0.05
Lymphocytes (10^3/g)	12.84(0-40.65)^a	23.75(1.12-66.3)^a	14.08(0-46.22)^a	>0.05
Neutrophils (10^3/g)	1090(484.54-1788.33)^b	2333.6(814.1-12390.5)^a	814.05(486.85-2726.72)^b	<0.05
Eosinophils (10^3/g)	13.16(0-68.89)^b	44.27(0-511.5)^a	22(0-209.46)^b	<0.001
Epithelial cells (10^3/g)	44.1(5.66-142.29)^a	165.44(36.15-302.1)^a	67.52(13.7-252.84)^a	>0.05

[2] BMI(Body Mass Index); OCS(Oral Corticosteroids); FENO(Fractional Exhaled Nitric Oxide); FEV1(Forced Expiratory Volume in one second); FVC(Forced Vital Capacity); TLC(Total Lung Capacity); RV(Residual Volume); DLCO(Diffusing Capacity for Carbon Monoxide); FRC(Functional residual capacity); LABA(Long Acting B2 Agonist); LAMA(Long Acting Muscarinic Antagonist); LTRA(Leukotriene Receptor Antagonist); CRP(C-Reactive Protein); CAT(COPD Assessment Test), ICS(Inhale Corticosteroids)

Cluster 2 and 3 had also more impaired lung function with more severe airway obstruction, lung hyper-distension, and severely reduced diffusing capacity and transfer coefficient (Figure 5.4). As far as inflammation is concerned, cluster 2 had more severe systemic and airway neutrophilic inflammation, with slightly raised fibrinogen but not CRP. Circulating lymphocytes were reduced in cluster 2 (Figure 5.5). Total Cell Counts ($10^6/g$) and Neutrophils (%) are presented in Figure 5.6. Absolute sputum eosinophil counts were higher in cluster 2 than in cluster 1 while no difference was seen in the blood (Figure 5.7). FeNO levels were similar between clusters and no difference was seen regarding total serum IgE (Figure 5.7) nor sensitizations to aeroallergens, which were low in all three clusters.

Table 5.6: Characteristics of patients with COPD after imputation, Median (IQR) / Percentage (frequency) in each cluster, and comparison between clusters

Result for clustering after multiple imputation				
Variable	**Cluster 1 (n=67)**	**Cluster 2 (n=56)**	**Cluster 3 (n=55)**	**P-value**
Demographic characteristics				
Age (Year)	62(55-66)^a	67(60.75-74.25)^b	67(58.5-73)^b	<0.001
Sex (Male)	76.12%(51)^a	75%(42)^a	7.27%(4)^b	<0.0001
Height (cm)	173(166.5-178)^a	169.5(162-17.25)^a	161(156-165)^b	<0.0001
Weight (kg)	77(65-89)^a	67(56.75-76)^b	60(53-66)^c	<0.0001
BMI (kg/m^2)	25.59(22.34-30.21)^a	23.08(20.55-24.89)^b	23.15(19.89-25.55)^b	<0.001
Cigarette Packs (Year)	36.9(21.52-50)^a	42.5(24.62-52.5)^a	34.5(21.42-43.87)^a	>0.05
Cigarettes (day)	20(10-25)^a	20(10-21.25)^a	20(10-20)^a	>0.05
Smoking Duration(Year)	40(31-46)^a	46.5(34.25-52.25)^b	44(37-50)^b	<0.05
OCS Course				<0.0001
0	80.59%(54)	30.35%(17)	61.82%(34)	
1	13.43%(9)	32.14%(18)	23.64%(13)	
≥ 2	4.48%(3)	35.71%(20)	10.91%(6)	
Antibiotic Course				>0.05
0	43.28%(29)	32.14%(18)	32.73%(18)	
1	46.27%(31)	50%(28)	58.18%(32)	
≥ 2	8.95%(6)	16.07%(9)	10.91%(6)	
Emergency Room Admission for asthma or COPD				<0.0001
0	95.52%(64)	71.43%(40)	85.45%(47)	
1	4.48%(3)	26.78%(15)	14.54%(8)	
≥ 2	0%(0)	1.78%(1)	0%(0)	
Number of hospitalizations for asthma or COPD				<0.0001
0	97.01%(65)	67.86%(38)	89.09%(49)	
1	2.98%(2)	26.78%(15)	10.91%(6)	
≥ 2	0%(0)	5.36%(3)	0%(0)	
Pulmonary characteristics				
FeNO (ppb)	15(10.5-23)^a	18.5(10-30.5)^a	14(8-25.5)^a	>0.05
FEV (mL)	1950(1700-2205)^a	1085(780-1352.5)^b	1240(1075-1390)^b	<0.0001
FEV1 (% predicted)	66(56-78)^a	37(30-45.25)^b	55(47.5-66)^c	<0.0001
FEV1 PD (mL)	2000(1820-2405)^a	1190(837.5-1380)^b	1310(1135-1470)^b	<0.0001
FEV1 PD (% predicted)	69(60.5-81.5)^a	40(32.75-50)^b	59(51.5-69.5)^c	<0.0001
Reversibility (%)	7(1-10.5)^a	6.5(1-11.25)^a	8(2-13.5)^a	>0.05
FVC (mL)	3200(2625-3780)^a	2427.5(1705-2767.5)^b	2140(1820-2390)^c	<0.0001
FVC (% predicted)	86(74-99.5)^a	65(55.75-74.25)^b	77(68.5-85.5)^c	<0.0001
FVC post (mL)	3320(2805-3870)^a	2385(1797.5-2900)^b	2130(1870-2445)^c	<0.0001
FVC post (%)	90(79.5-101)^a	68(56.75-76.5)^b	80(71.5-91.5)^c	<0.0001
FEV1/ FVC pre (%)	63.1(56.45-66.65)^a	47(40.37-50.42)^b	57.1(53.15-64.5)^a	<0.0001
FEV1/ FVC post (%)	64(58.3-68.55)^a	47.5(42.8-53.2)^b	59.9(54.5-67.05)^a	<0.0001
TLC (mL)	6725(5725-7385)^a	6830(5847.5-8107.5)^a	5440(4975-6005)^b	<0.0001
TLC (% predicted)	103(93-112)^a	113(100.75-125.25)^b	115(102-127.5)^b	<0.0001
RV (mL)	3320(2970-3820)^a	4515(3690-5277.5)^b	3270(2920-3725)^a	<0.0001
RV (% predicted)	141(128-171.5)^a	183.5(163.87-228.5)^b	172(148.5-193.25)^c	<0.0001
RV/TLC (%)	52.99(46.71-57.18)^a	66.37(61.28-69.14)^b	60.84(57.03-64.06)^c	<0.0001
DLCO (mmol/kPa.min)	5.2(4.06-5.88)^a	3.47(2.67-4.44)^b	3.27(2.69-4.03)^b	<0.0001
DLCO (% predicted)	57(46.5-67)^a	41.5(30.75-58)^b	45(37.5-53.5)^b	<0.0001
DLCO/AV (mmol/kPa.min/l)	1.03(0.79-1.28)^a	0.80(0.63-0.99)^b	0.93(0.71-1.15)^b	<0.0001
DLCO/AV (% predicted)	70(57-88)^a	59.5(45.75-79)^b	60(50-77.5)^b	<0.0001
sGaw (1/kPa*sec)	0.65(0.44-0.88)^a	0.35(0.24-0.45)^b	0.47(0.34-0.73)^c	<0.0001
FRC (L)	4.83(4.08-5.43)^a	5.87(4.82-6.49)^b	4.24(3.77-4.95)^c	<0.0001
FRC (% predicted)	143(122-161)^a	174(154.75-192)^b	160.5(141.5-180.5)^c	<0.0001
CAT score	22(15-31)^a	26(19-30)^a	24(17-30)^a	>0.05
Treatment characteristics				
Treatment (Yes)	44.78%(30)^a	78.57%(44)^b	67.27%(37)^b	<0.0001
ICS (Yes)	29.85%(20)^a	71.43%(40)^b	69.10%(38)^b	<0.0001
OCS (Yes)	1.49%(1)^a	16.07%(9)^b	0%(0)^a	<0.0001
LAMA (Yes)	34.33%(23)^a	73.21%(41)^b	49.09%(27)^a	<0.0001
LABA (Yes)	44.78%(30)^a	82.14%(46)^b	76.36%(42)^a	<0.0001
SABA (Yes)	25.37%(17)^a	50%(28)^b	49.09%(27)^b	<0.001
LTRA (Yes)	2.98%(2)^a	0%(0)^a	7.27%(4)^a	>0.05
Theophylline (Yes)	1.49%(1)^a	5.36%(3)^a	1.82%(1)^a	>0.05

Table 5.6 – continued from previous page

Result for clustering after multiple imputation				
Variable	**Cluster 1 (n=67)**	**Cluster 2 (n=56)**	**Cluster 3 (n=55)**	**P-value**
Blood characteristics				
Leucocytes (1/μl)	7.71(6.52-9.43)^a	8.56(6.87-11.10)^a	7.81(6.64-9.2)^a	>0.05
Neutrophils (%)	58.2(52.75-63.55)^a	67.6(60.1-73.22)^b	58.3(53.25-64.85)^a	<0.0001
Lymphocytes (%)	30.3(25.65-35)^a	20.95(15.7-27.27)^b	29(23.82-35.12)^a	<0.0001
Monocytes (%)	8.7(7.1-9.8)^a	7.8(6.34-9.82)^b	7.1(6.4-8.67)^b	<0.05
Eosinophils (%)	2(1.1-3.25)^a	1.9(0.9-3.22)^a	2.1(0.95-3.35)^a	>0.05
Basophils (%)	0.4(0.3-0.6)^a	0.3(0.2-0.5)^a	0.4(0.3-0.7)^a	>0.05
Neutrophils (1/μl)	4529.13(3723.5-5501.7)^a	5574.91(4097.7-7842.7)^b	4840.71(3393.7-5743.1)^a	<0.05
Lymphocytes (1/μl)	2346.8(1844.6-2968.1)^a	1847.22(1340.4-2382.2)^b	2253.3(1846.6-2710.8)^a	<0.0001
Monocytes (1/μl)	668.82(507.24-831.28)^a	652.58(511.72-864.82)^a	605.82(485.32-748.05)^a	>0.05
Eosinophils (1/μl)	138(75.12-240.24)^a	168.78(69.23-258.42)^a	147.63(88.26-255.97)^a	>0.05
Basophils (1/μl)	30.84(21.48-42.58)^a	27.63(18.88-46.25)^a	32.04(23.95-48.97)^a	>0.05
Fibrinogen (g/l)	3.43(2.88-3.86)^a	3.65(3.08-4.50)^b	3.54(2.98-3.80)^b	<0.05
CRP (mg/l)	2.3(1.25-4.95)^a	2.9(0.87-7.7)^a	1.8(1-4.75)^a	>0.05
Alpha 1 antitrypsin (g/l)	1.49(1.34-1.57)^a	1.61(1.39-1.76)^b	1.49(1.35-1.64)^a	<0.05
Calcium (mmol/L)	2.42(2.37-2.47)^a	2.41(2.35-2.48)^a	2.44(2.38-2.47)^a	>0.05
25(OH) Vitamine D (ng/ml)	18(12-30.9)^a	17.5(10-26.75)^a	24(15.5-33)^a	>0.05
Phosphate (mmol/L)	0.89(0.77-0.95)^a	0.88(0.72-1.06)^a	1.01(0.93-1.11)^b	<0.0001
Atopy characteristics				
IgE (KU/L)	52(23-184.5)^a	91.5(28.75-304.75)^a	73(21-170)^a	>0.05
RAST DPT (d1) %>0.35 (KU/L)	13.43%(9)^a	16.07%(9)^a	7.27%(4)^a	>0.05
RAST Cat (e1), %>0.35 (KU/L)	1.49%(1)^a	7.14%(4)^a	1.82%(1)^a	>0.05
RAST Dog (e5), %>0.35 (KU/L)	2.98%(2)^a	5.36%(3)^a	0%(0)^a	>0.05
RAST Grass (GX3), %>0.35 (KU/L)	5.97%(4)^a	14.28%(8)^a	1.82%(1)^a	>0.05
RAST microog (MIX1), %>0.35 (KU/L)	7.46%(5)^a	12.5%(7)^a	10.91%(6)^a	>0.05
RAST Birch (t3), %>0.35 (KU/L)	0%(0)^a	3.57%(2)^a	0%(0)^a	>0.05
Sputum characteristics				
Positive Aerobic Sputum Culture	8.95%(6)^a	16.07%(9)^a	12.73%(7)^a	>0.05
Weight of sputum (g)	1.77(1.28-3.01)^a	1.45(0.97-3.05)^a	1.65(0.93-3.28)^a	>0.05
Total Cell Counts (10^6/g)	1.70(0.90-4.41)^a	4.90(1.58-15.96)^b	2.65(1.08-5.43)^a	<0.05
Squamous (%)	19(6-42)^a	7(2-31.12)^b	9.5(2-23.5)^b	<0.05
Viability (%)	68(53.5-85.5)^a	69(51-86)^a	67(48.5-79.5)^a	>0.05
Macrophages (%)	14(6.7-26.1)^a	6.5(2.54-15.15)^b	17(7-29.95)^a	<0.0001
Lymphocytes (%)	1.2(0-2.3)^a	0.6(0.15-1.5)^a	1.2(0-3)^a	>0.05
Neutrophils (%)	71.4(56.35-88.2)^a	78.9(58.8-92.19)^a	60.4(39.3-81.05)^b	<0.001
Eosinophils (%)	1.4(0.2-4.92)^a	2.3(0.2-9.12)^a	2(0.7-9.15)^a	>0.05
Epithelial cells (%)	2.4(0.4-8.1)^a	3.25(0.57-11.05)^a	4.7(1.3-15.2)^a	>0.05
Macrophages (10^3/g)	233.55(53.28-691.65)^a	301.29(31.37-755.45)^a	330.48(58.13-816.67)^a	>0.05
Lymphocytes (10^3/g)	14.2(0-43.07)^a	23.75(4.39-63.9)^a	25.2(3.42-72.5)^a	>0.05
Neutrophils (10^3/g)	1090(490.86-2339.64)^a	2482(884-117218)^b	1110(493.04-3483.82)^a	<0.05
Eosinophils (10^3/g)	19.92(0-166.96)^a	71.67(0-638.26)^b	33.04(5.65-293.7)^a	<0.001
Epithelial cells (10^3/g)	38.45(5.66-142.29)^a	94(18.07-292.65)^a	81.36(16.4-301.24)^a	>0.05

[3] BMI(Body Mass Index); OCS(Oral Corticosteroids); FENO(Fractional Exhaled Nitric Oxide); FEV1(Forced Expiratory Volume in one second); FVC(Forced Vital Capacity); TLC(Total Lung Capacity); RV(Residual Volume); DLCO(Diffusing Capacity for Carbon Monoxide); FRC(Functional residual capacity); LABA(Long Acting B2 Agonist); LAMA(Long Acting Muscarinic Antagonist); LTRA(Leukotriene Receptor Antagonist); CRP(C-Reactive Protein); CAT(COPD Assessment Test), ICS(Inhaled Corticosteroids)

5.4 Discussion

In this chapter, we characterized COPD patients into three distinctly different clusters by applying general and flexible statistical computation in the dataset with missing values. In the present chapter, clustering was applied to a large number of variables instead of selecting a limited number of variables. Although missing values are a pervasive problem in diverse datasets such as COPD with large number of variables, missing values have not been considered properly in the clustering literature. Based on these restrictions, the classification of COPD datasets has not been comprehensively investigated. Therefore, in this exhaustive study, phenotypes in COPD dataset were described by imputing missing values, principal components, and cluster anal-

ysis with many analytical decisions, which overcome limitations, often reported in previous clustering studies.

The concept of treatable trait has become very popular over the last years and it has been suggested to avoid the label of asthma or COPD among patients with severe chronic airway diseases (Agusti et al., 2016). Adopting this taxonomic view COPD population could be seen as a population featuring the trait of fixed airway obstruction after a significant smoking history and denying any previous diagnosis of asthma before the age of 40. One strength of this study, compared to previous clustering analysis in COPD, is that it included airway inflammatory features, FeNO, and atopic status in the parameters subjected to analysis.

We actually found 3 clusters of COPD, strikingly linked to sex with two clusters showing male dominance while the third was essentially a female cluster. There were clear differences between lung function impairment between the clusters whereas quantitative smoking history was quite similar, pointing to different susceptibility to tobacco among patients. The percentage of contribution of the different variables to the clustering. It appears that functional criteria including airway flow and the degree of airway obstruction and lung hyper-distension and the % of blood lymphocytes are amongst the most important criteria to structure the cluster while variables like smoking history, FeNO, and Vitamin D level were rather homogeneous between the patients.

Cluster 2 is conspicuously the most severe group of patients with marked airway obstruction, lung hyper-distension capacity together with intense neutrophilic inflammation both at the systemic and the airway level, in keeping with the previously reported relationship between the severity of airway obstruction and the neutrophilic inflammation (Moermans et al., 2011). Cluster 2 and 3 had also impaired diffusing capacity and transfer coefficient pointing to emphysema. Despite severe emphysema the level of α_1-antitrypsin was higher in cluster 2, perhaps indicating a response of the body trying to counteract the lung destruction favoured by the intense neutrophilic inflammation.

Associated with neutrophilic inflammation, cluster 2, displays a small rise in fibrinogen level even through the median value remained within the normal range and below the threshold of 5.1 g/l, shown to be predictive of an excess of mortality (Celli et al., 2012). Of note is the fact that the intensity in neutrophilic airway inflammation

is not associated with bacterial colonization identified by classical bacterial culture, which was rather low come close to 10% for the whole cohort. Of course, it does not imply that the microbiome may be profoundly disturbed in COPD and more sophisticated microbiological analyses might have revealed differences between the clusters. The altered microbiome may be the consequence of frequent antibiotic courses received by the patients as shown in the cohort since almost two-thirds of the patients had received antibiotics for bronchitis the year prior to the visit. It is worth noting that there was no difference between clusters in the number of antibiotic courses. As opposed to exacerbation defined by OCS course, exacerbation defined by the antibiotic course was not related to the severity of lung function impairment nor to the severity of airway inflammation.

We also looked at phosphocalcic metabolism and found, as expected in western Europe, reduced levels of 25 OH Vit D (< 30 ng/ml) in all clusters without any difference between them. While Calcium levels, a tightly regulated ion, were normal and similar between the three clusters, there were striking differences in the levels of phosphates, which were clearly lower in cluster 1 and also cluster 2. The literature about phosphate level in COPD is virtually absent and the clinical meaning of our finding remains obscure though there might be a sex effect as this demographic trait best differentiates cluster 1 and 2 from cluster 3.

The eosinophilic trait is a marker of response to corticoids in asthma but also in COPD. In this study, blood eosinophilic inflammation does not appear to be a discriminant feature between the clusters but cluster 2 shows a greater absolute, but not relative, sputum eosinophil cell count. However, the three clusters had median value of sputum eosinophil counts greater than that we found in a healthy population (Demarche et al., 2016) and, actually, rather similar to what is found in large population of unselected asthmatics (Schleich et al., 2014). Furthermore, the greater amount of eosinophils present in the sputum of COPD was noted despite heavier treatment with ICS in this cluster, which points to some corticoresistance in this cluster. Atopic status based on positive RAST towards aeroallergens was low in the three clusters but total serum IgE was measured at a higher level than those usually found in a general population of this age, maybe pointing to a role of IgE mediated processes in the pathophysiology of the disease. It has also been suggested that smoking may stimulate IgE production (O'Connor et al., 1989). However, it is a production clearly directed towards something different from the classical aeroallergens encountered in asthma. The reason why atopic status is so low in COPD, clearly lower than in a

general population, is likely to be related to the age of the COPD patients. Indeed, it was demonstrated in population studies that specific IgE levels decreased with age (Amaral et al., 2016) and we have shown in a large asthmatic population that the rate of sensitization to aeroallergens was sharply declining with age after 60 years (Manise et al., 2016).

Of interest, and also perhaps surprising, is the fact the CAT score does not differ between the clusters despite clear differences in lung function impairment and extent in airway inflammation. This shows that CAT score cannot capture the airflow limitation nor airway inflammation, indicating that symptoms, lung function, and airway inflammation are different domains accounting for the disease variability (Lapperre et al., 2004).

This study has obviously limitations as we have not taken into account comorbidities, exercise capacity, and lung imaging in our parameters, which are key variables in phenotyping the COPD patients in clinical practice. Our purpose here was rather to explore variables traditionally linked to asthma such as FeNO, IgE, and eosinophilic inflammation, and to see whether they can play a significant contribution in defining the variability of the disease in patients > 40 years with smoking history and persistent airway obstruction. Our data indicate that FeNO, blood eosinophils, and serum IgE, though being significantly different from what is found in a healthy population for total serum IgE, are not able to single out a particular cluster. Therefore, most of the T2 biomarkers had not enough variability among the patients to shape a cluster. However, eosinophilic (together with neutrophilic) airway inflammation is raised in the cluster that shows the most severe lung function impairment. These data may have importance as it has been shown by retrospective post hoc analysis that ICS treatment in eosinophilic COPD might actually slow down the lung function decline (Pavord et al., 2016). The impact of targeting eosinophilic inflammation in COPD should be given careful consideration in long-term clinical trials using not only ICS but also anti-interleukine-5.

Another important limitation of this study is the lack of a validation cohort whereas Castaldi et al. (2017) has shown that the reproducibility of COPD clustering across studies was rather modest. The size of our cohort was however too small to split our population and perform meaningful clustering with our extensive set of variables. However, we presented validation with two statistical indices.

In conclusion, in a cohort of COPD, we found 3 clusters of patients with similar age and smoking history but very different sex distribution and lung function, and inflammatory parameters. In particular, we identified a cluster of male patients with intense granulocytic airway inflammation combined with severe airway flow limitation and lung hyperdistention, who are prone to exacerbate and undergo recurrent hospitalizations. These clusters need to be confirmed in a new cohort of patients, ideally from other centers.

5.5 Appendix

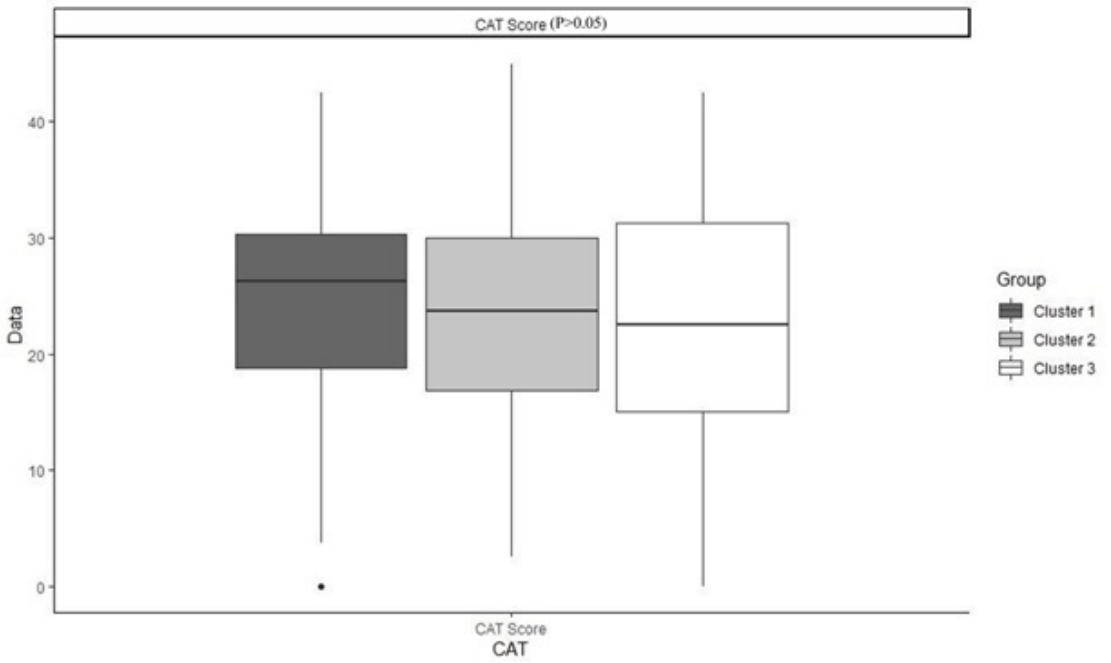

Figure 5.3: CAT Score in three clusters

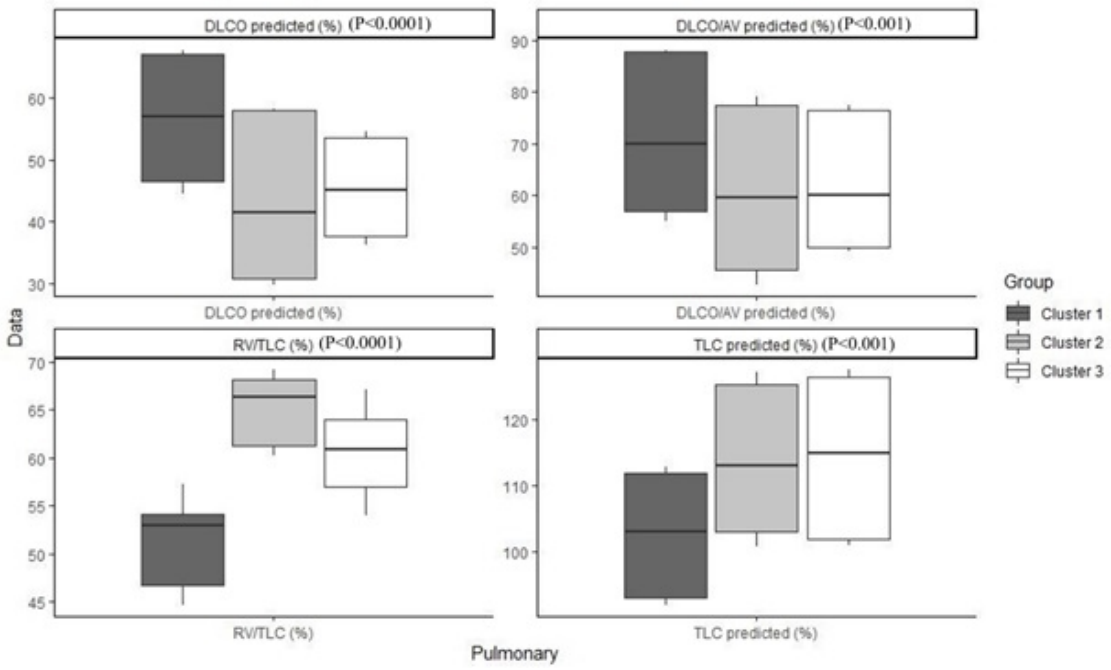

Figure 5.4: Lung function in three clusters

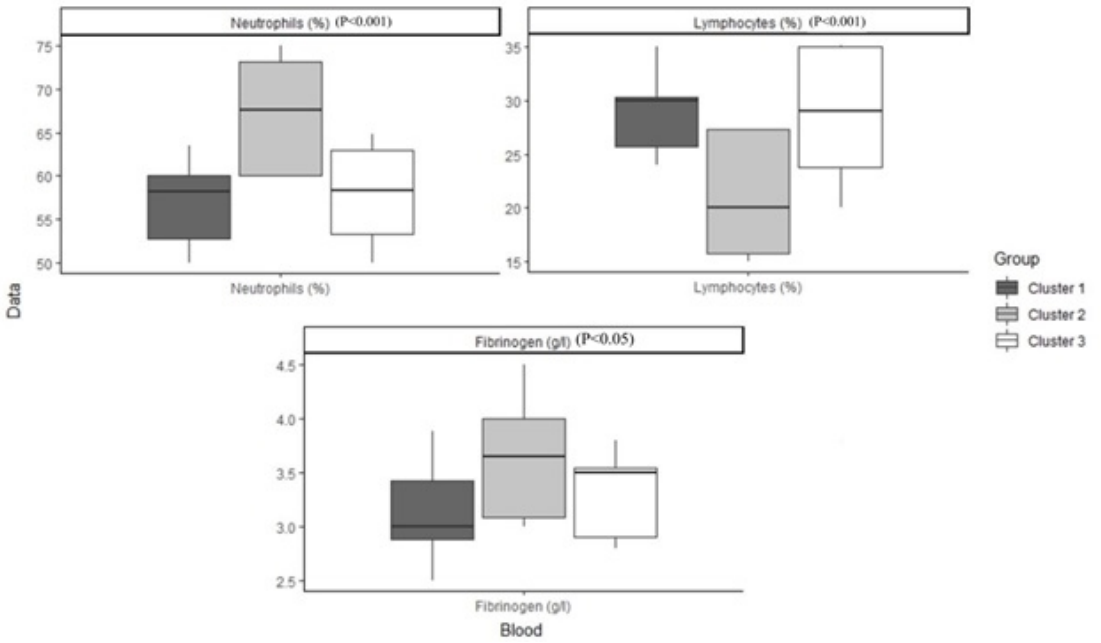

Figure 5.5: Blood in three clusters

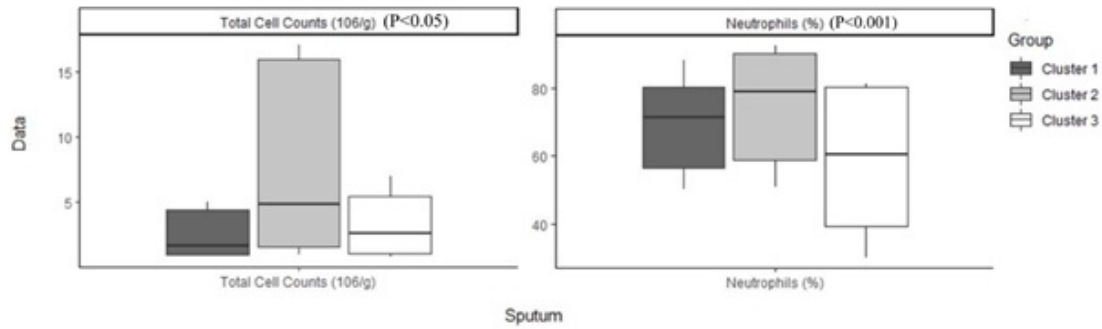

Figure 5.6: Sputum in three clusters

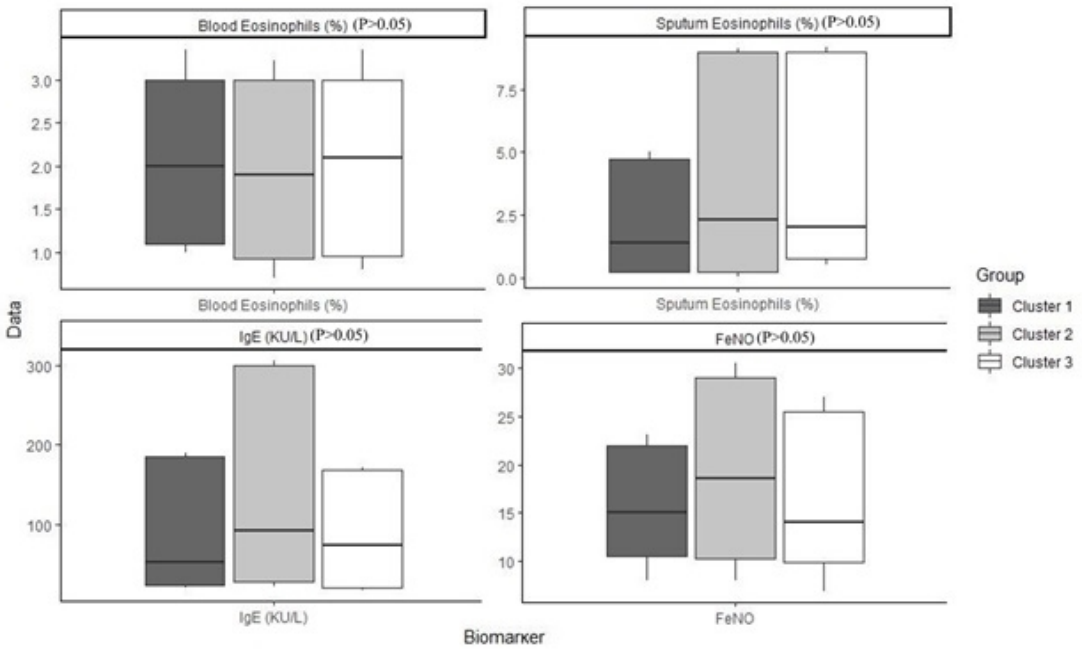

Figure 5.7: T2 biomarkers in three clusters

CHAPTER 6

Clustering on the eosinophilic asthmatic patients

As explained in Chapter 1, eosinophilic asthma is a phenotype that, despite being overall associated with disease severity, may sometimes be present in milder disease. Although eosinophilic asthma is known as a main phenotype of asthma classification, in this thesis, we recognized heterogeneity in this phenotype and attempt to identify groups of patients presenting similar characteristics. Therefore, the main purpose of this chapter is to perform a cluster analysis on a large group ($n = 426$) of eosinophilic (sputum eosinophils $\geq 3\%$) asthmatics. There is a possibility that ICS treatment at various doses could be a confounding factor. As a result, classification is reapplied to steroid naïve and high dose ICS treated patients. Section 6.1 presents the characteristics of eosinophilic asthmatic patients as well as those of the two subsets of ICS with the same variables, along with a comprehensive description of missing values. The proposed clustering framework is presented in Section 6.2. The results of clustering for the whole eosinophilic asthmatic cohort, steroid naïve cohort, and high dose ICS treated cohort are presented in Section 6.3. In Section 6.4, clinical interpretations and explanations of the results of clustering are provided.

This Chapter is based on

Nekoee Zahraei, H., Guissard, F., Paulus, V., Henket, M., Donneau, A.F., & Louis, R., Clustering on the eosinophilic asthmatic patients. (manuscript in under-review)

6.1 Eosinophilic asthmatic dataset

Eosinophilic airway inflammation is a major trait of asthma (Agusti et al., 2016). It is accepted that a sputum eosinophil count reaching 2−3% is considered as a sign of a significant eosinophilic inflammation (Brightling, 2006). Large cross-sectional studies have shown that a sputum eosinophil count of at least 3% is found in almost 50% of asthmatics seen in a secondary care center (Schleich et al., 2013) and this proportion can further increase to more than 60% when severe patients are selected (Graff et al., 2020). Overall, patients with sputum eosinophils above 2 − 3% display a more severe disease with poorer asthma control and greater health care utilization as compared to their non-eosinophilic counterparts (Demarche et al., 2017; Hastie et al., 2021). In patients with steroid naïve asthma, the eosinophilic trait generally predicts good clinical response to ICS (Brightling, 2006) whereas some patients treated with high dose ICS, and possibly OCS, may still show severe eosinophilic inflammation associated with poor clinical outcomes (Louis and Schleich, 2021; Graff et al., 2020; van Bragt et al., 2020) pointing out a disease in which the eosinophilic inflammation is relatively resistant to corticoids.

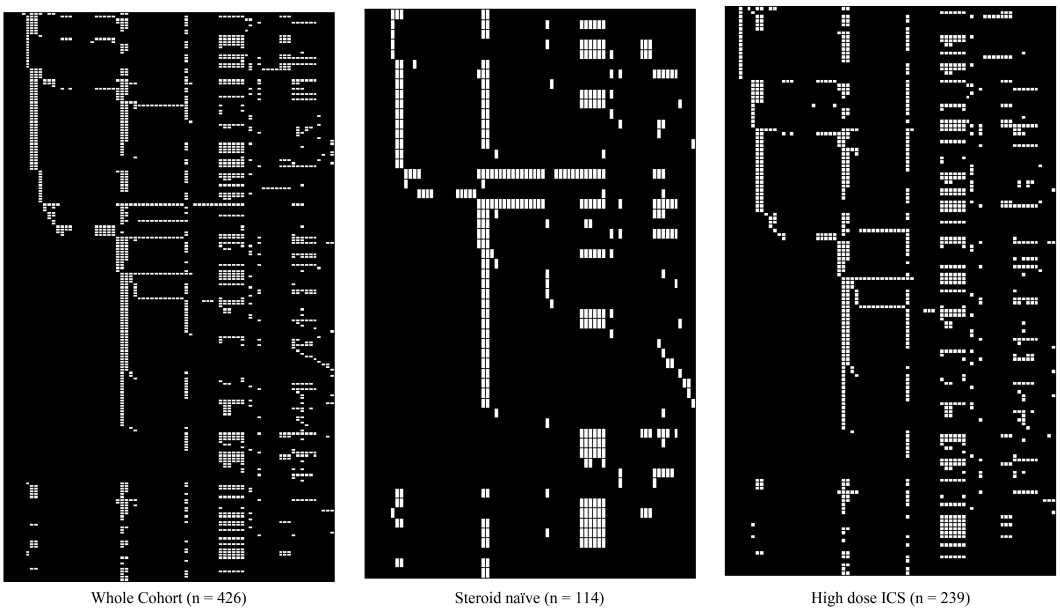

Figure 6.1: A general overview of eosinophilic asthmatic cohorts (white cells are missing values)

While eosinophils are usually thought to be potent inflammatory cells and an active contributor to asthma severity some authors have suggested that lung tissue may actually harbor a population of regulatory eosinophils the function of which might be

to dampen airway inflammation (Mesnil et al., 2016). A recent study has highlighted the existence of a group of patients with mild asthma and high sputum eosinophil count (Xu et al., 2021) questioning the detrimental action of eosinophils in shaping disease severity and perhaps suggesting heterogeneity in their functional roles. This chapter examined a large dataset of asthma clinic patients with a sputum eosinophil ≥ 3% which was recruited from the asthma clinic of Liege University Hospital between 2011 and 2020.

As part of this chapter, the whole eosinophilic asthmatic cohort was divided into two subgroups: patients without ICS (steroid naïve cohort), and patients with ICS >1000g/d equivalent beclomethasone (high dose ICS treated cohort). Qualitative variables were summarized as count and percentage, while median and interquartile range (P25 - P75) were calculated as quantitative variables, in Tables 6.1. In addition, the count and percentage of missing values were included in these tables. In Figure 6.1, a general perspective of the whole eosinophilic asthmatic cohort, and two subgroups were presented. Variables are represented by columns, and patients by rows. Missing data is represented by white cells.

According to Table 6.1, patients were mostly females (55%) with a median age of 53, and 56% were atopic. Patients displayed a mild overweight (median body mass index was 26). The median asthma duration was 15 years. There are 56% of patients who have not experienced an exacerbation the year prior to the visit, 15% have had it once, and 18% have had it more than once. The missing values were MCARs. The percentage of missing values ranged from 0% to 76% for DHEA sulfate (μmol/L) and 96% of patients presented at least one missing value.

There were 114 steroid naïve eosinophilic asthmatic patients with an equal number of males and females, a majority of non-smokers (78%), and atopic (60%) with a median age of 50 and age at diagnosis of 30 years. The median BMI was 26. 65% of patients had not experienced an exacerbation the year prior to the visit, 14% had experienced it once, and 14% had experienced it more than once.

Table 6.1: Descriptive statistics of characteristics of eosinophilic asthmatic cohort

Variable	Whole cohort (n = 426)		Steroid naïve (n = 114)		High dose ICS (n = 239)	
	Median (IQR) %(n)	Missing Value %(n)	Median (IQR) %(n)	Missing Value %(n)	Median (IQR) %(n)	Missing Value %(n)
Demographic characteristics						
Age (Year)	53(40–64)	0%(0)	50(36–64)	0%(0)	53(40–64)	0%(0)
Sex (Male)	45% (190)	0% (0)	50%(57)	0%(0)	41%(97)	0%(0)
BMI (kg/m^2)	26 (23 - 29)	0.2% (1)	26(23–28)	0%(0)	26(23–30)	0.4%(1)
Smoking		0.7% (3)		0%(0)		1%(3)
Ex-smoker	30% (126)		3%(26)		33%(79)	
Smoker	15% (62)		19%(22)		13%(31)	
Duration of smoking(Year)	0 (0 - 14)	8.2% (35)	0(0–10)	7%(8)	0(0–14)	8%(20)
Age at diagnosis (Year)	30 (10 - 51)	21% (88)	30(10–57)	24%(27)	29(10–50)	18%(45)
Duration of asthma(Year)	15 (3 - 30)	21% (88)	6(0–24)	24%(27)	18(8–32)	19%(45)
Atopy (Yes)	56%(238)	7% (30)	60%(69)	5%(6)	56%(133)	7%(18)
Comorbidities						
Nasal polyposis (Yes)	11% (45)	1% (4)	6%(7)	1%(1)	12%(29)	1%(2)
Allergic rhinitis (Yes)	30% (124)	2% (7)	32%(36)	2%(2)	29%(70)	2%(4)
GERD (Yes)	13% (55)	4% (16)	12%(14)	2%(2)	13%(32)	4%(9)
Treatment						
ICS (Yes)	69%(294)	4%(18)				
ICS (μg/d)	1000 (0 - 2000)	4%(18)				
LABA (Yes)	65% (277)	7% (32)	95%(3)	5%(6)	90%(215)	8%(20)
LTRA (Yes)	25% (107)	0.5% (2)	7%(8)	0%(0)	38%(90)	1%(2)
LAMA (Yes)	6% (26)	7% (32)	2%(2)	5%(6)	9%(22)	8%(20)
Theophylline (Yes)	1.2% (5)	0.5% (2)	0%(0)	0%(0)	2%(4)	1%(2)
H_1 antagonists (Yes)	20% (86)	0.5% (2)	13%(15)	0%(0)	24%(57)	1%(2)
SABA.SAMA (Yes)	65% (279)	7% (32)	62%(71)	5%(6)	69%(164)	8%(20)
OCS (Yes)	11% (49)	0.5% (2)	3%(4)	3%(4)	18%(42)	1%(2)
Asthma control, exacerbation, and quality of life						
ACT	16 (11 - 21)	1% (5)	18(15–22)	2%(2)	13(9–19)	0.8%(2)
ACQ	1.8 (1 – 2.9)	2% (7)	1(0.7–2)	1%(1)	2.6(1.4–3.4)	2%(5)
Number of exacerbation		9% (39)		7%(8)		11%(28)
0	56% (239)		65%(74)		50%(119)	
1	18% (78)		14%(16)		20%(48)	
≥ 2	16% (70)		14%(16)		18%(44)	
AQLQ Global	4.6 (3.4 - 5.8)	2% (10)	5(4–6)	2%(2)	4(2.9–5.3)	2.5%(6)
Pulmonary function						
FEV1 pre (% predicted)	85(69 - 98)	0.2% (1)	93(82–102)	1%(1)	75(62–92)	0%(0)
FEV1 post (%predicted)	90(77 - 103)	0.2% (1)	100(91–110)	1%(1)	69(83–97)	0%(0)
FVC pre (% predicted)	96(85 - 107.25)	0.5% (2)	105(93–113)	1%(1)	88(77–100)	0.4%(1)
FVC post (% predicted)	93(81 - 106)	0.5% (2)	102(91–112)	1%(1)	92(80–102)	0.4%(1)
FEV1/ FVC pre (%)	73(66 - 80)	0.2% (1)	75(70–81)	1%(1)	71(62–79)	0%(0)
FEV1/ FVC post (%)	77(69 - 83)	0.5% (2)	80(74–84)	1%(1)	74(66–82)	0%(0)
FRC (% predicted)	118(98 – 138)	31% (132)	110(94–133)	25%(29)	122(102–140)	34%(81)
DLCO/VA (% predicted)	95(83 - 108)	31% (132)	95(82–111)	27%(31)	93(83–106)	34%(81)
DLCO (% predicted)	79(69 - 92)	31% (132)	82(73–92)	27%(31)	76(68–90)	34%(81)
RV (% predicted)	116(92 - 141)	28% (119)	101(83–128)	25%(29)	127(102–156)	29%(70)
TLC (% predicted)	98(89 - 107)	28% (119)	97(88–106)	25%(29)	99(89–107)	29%(70)
sGaw (1/kPa*sec)	0.8(0.6 - 1.1)	33% (139)	1(0.7–1.2)	28%(32)	0.7(0.4–1)	34%(82)
PC20M(mg/ml)	1.8(0.6 - 8)	46% (194)	1.8(0.6–7)	10%(12)	1.6(0.6–8.5)	64%(153)
FeNO (ppb)	37(19 - 62)	1% (4)	41(21–76)	2%(2)	35(19–54)	0.8%(2)
Sputum cell count						
Weight of sputum (g)	2.2(1.2 - 3.9)	1% (4)	2(1–4)	2%(2)	2.3(1.2–3.9)	0.4%(1)
Total cell counts (10^6/g)	1.7(0.9 - 4.5)	2% (7)	1.5(1–5)	1%(1)	1.8(0.8–4.5)	1.7%(4)
Viability (%)	70(56 - 81)	1% (4)	71(51–80)	1%(1)	70(57–82)	0.4%(1)
Squamous (%)	12(4 – 27)	1% (4)	14(7–30)	1%(1)	11(4–27)	0.8%(2)
Macrophages (%)	21(10 - 36)	0% (0)	26(14–37)	0%(0)	18(8–35)	0%(0)
Lymphocytes (%)	1(0.4 - 2)	0% (0)	1(0.6–3)	0%(0)	0.8(0.2–1.8)	0%(0)
Neutrophils (%)	49(31 - 69)	0% (0)	50(32–68)	0%(0)	50(30–69)	0%(0)
Eosinophils (%)	11(5 - 26)	0% (0)	10(5–21)	0%(0)	11(5–30)	0%(0)
Epithelial cells (%)	3(1 - 7)	0.2% (1)	3(1–5)	0%(0)	3(2–8)	0.4%(1)
Macrophages (10^3/g)	363(140 - 866)	1.6% (7)	433(149–1260)	1%(1)	325(123–787)	1.7%(4)
Lymphocytes (10^3/g)	17(3 - 47)	1.6% (7)	26(5–86)	1%(1)	13(2–39)	1.7%(4)
Neutrophils (10^3/g)	702(270 - 2124)	1.6% (7)	715(253–2746)	1%(1)	660(267–2120)	1.7%(4)
Eosinophils (10^3/g)	211(73 - 711)	1.6% (7)	229(65–711)	1%(1)	214(68–784)	1.7%(4)
Epithelial cells (10^3/g)	56(17 - 147)	1.9% (8)	43(17–139)	1%(1)	67(16–147)	2.1%(5)

Table 6.1 – continued from previous page

Variable	Whole cohort		Steroid naïve (n = 114)		High dose ICS (n = 239)	
	Median (IQR) %(n)	Missing Value %(n)	Median (IQR) %(n)	Missing Value %(n)	Median (IQR) %(n)	Missing Value %(n)
Blood cell count						
Leucocytes ($10^3/\mu$l)	7(6 - 9)	1.4% (6)	7(6–8)	2%(2)	8(7–9)	1.3%(3)
Neutrophils (%)	54(48 - 61)	1.4% (6)	52(46–58)	2%(2)	55(48–62)	1.3%(3)
Lymphocytes (%)	32(26 - 39)	1.4% (6)	35(30–40)	2%(2)	31(24–37)	1.3%(3)
Monocytes (%)	8(6 – 9)	1.4% (6)	8(6–9)	2%(2)	8(6–9)	1.3%(3)
Eosinophils (%)	4(2 – 6)	1.4% (6)	3(2–5)	2%(2)	4(3–6)	1.3%(3)
Basophils (%)	0.5(0.4 - 0.7)	1.4% (6)	0.5(0.4–0.7)	2%(2)	0.5(0.3–0.8)	1.3%(3)
Neutrophils ($1/\mu$l)	3942(3178 – 5171)	1.4% (6)	3648(2844–4475)	2%(2)	4201.5(3387–5435)	1.3%(3)
Lymphocytes ($1/\mu$l)	2350(1909 - 2872)	1.4% (6)	2368(1955–2868)	2%(2)	2359(1922–2871)	1.3%(3)
Monocytes ($1/\mu$l)	567(460 - 718)	1.4% (6)	539(428–672)	2%(2)	590(477–720)	1.3%(3)
Eosinophils ($1/\mu$l)	278(182 - 445)	1.4% (6)	231(178–344)	2%(2)	320(190–506)	1.3%(3)
Basophils ($1/\mu$l)	39(27 - 58)	1.4% (6)	37(24–52)	2%(2)	41(28–62)	1.3%(3)
Serum IgE						
RAST Birch (t3) %>0.35 (KU/L)	21% (89)	9.4% (40)	26%(77)	6%(7)	20%(48)	10%(24)
RAST Mould (MIX1) %>0.35 (KU/L)	15% (62)	8.9% (38)	10%(11)	9%(10)	18%(44)	9.2%(22)
RAST Grass (GX3) %>0.35 (KU/L)	31% (130)	11.3% (48)	33%(38)	11%(12)	31%(75)	12%(28)
RAST Dog (e5) %>0.35 (KU/L)	30% (127)	6.8% (29)	35%(40)	5%(6)	29%(70)	8%(19)
RAST Cat (e1) %>0.35 (KU/L)	29% (125)	6.3% (27)	35%(40)	4%(5)	29%(69)	7%(18)
RAST DPT (d1) %>0.35 (KU/L)	38% (162)	5.4% (23)	42% (48)	4%(5)	38%(90)	5%(13)
Total IgE(KU/L)	189(60 – 470)	6.3% (27)	201(61–396)	6%(7)	226(64–522)	7%(18)
Systematic inflammation						
CRP (mg/l)	2(1 - 5)	3% (13)	1.5(0.7–3.4)	3%(3)	2.1(1–5)	5%(10)
Fibrinogen (g/l)	3.35(2.83 - 3.79)	4.7% (20)	3.21(2.67–3.61)	5%(6)	3.38(2.84–3.87)	5%(11)
Adrenal function						
Cortisol (nmol/l)	207(162 - 271)	57% (244)	243(185–309)	78%(89)	203(155–264)	52%(125)
DHEA sulfate (μmol/L)	2.6(1.5 - 5.2)	76% (326)	2.6(1.6–5.6)	79%(90)	2.5(1.4–4.6)	76%(181)

[1] BMI(Body Mass Index); GERD(Gastroesophageal Reflux Disease); LABA(Long Acting B2 Agonist); LAMA(Long Acting Muscarinic Antagonist); LTRA(Leukotriene Receptor Antagonist); ICS(Inhaled Corticosteroids); OCS(Oral Corticosteroids); CRP(C-Reactive Protein); DHEAS(Dehydroepiandrosterone Sulfate); IgE(Immunoglobulin E); FENO(Fractional Exhaled Nitric Oxide); FEV1(Forced Expiratory Volume in one second); FVC(Forced Vital Capacity); TLC(Total Lung Capacity); RV(Residual Volume); DLCO(Diffusing Capacity for Carbon Monoxide); FRC(Functional residual capacity)

A subset of 239 patients with eosinophilic asthma received high-dose ICS. These patients were mostly female (59%) with a median age of 53 years, a median asthma duration was 15 years and 56% were atopic. The majority (54%) had never smoked. Patients displayed a mild overweight (median body mass index was 26). 50% of patients had not experienced an exacerbation the year prior to the visit, 20% had experienced it once, and 18% had experienced it more than once.

6.2 Clustering Framework

In order to account for the uncertainty of missing values in analysis, multiple imputation was applied for handling missing values. Multiple imputation generates a set of m (m = 100) independent plausible values for each missing value (Section 3.1.3). For the second step, FAMD was performed independently on m imputed dataset for reducing the complexity of huge dimensional data (Section 3.2.2). Then, hierarchical clustering was applied using Ward's criterion on the principal components derived in the first step. The number of clusters for each imputed dataset was determined using a package of 30 indices for determining the relevant number of clusters, and then K-means was applied for assigning clusters to patients according

to the number of clusters detected. Since 100 results have been recorded for clustering, in the final step, consensus clustering was considered based on 4M method (Section 3.3.3) to assign each individual to a cluster. The proposed framework is presented in Table 6.2.

Mann-Whitney nonparametric test for quantitative variables and chi-square for qualitative variables were applied on the original dataset to assess the comparison of patients'characteristics between clusters. Finally, the difference between the groups was depicted by boxplots for the quantitative variable. All analyses were performed using R statistical software with several well-known packages. *MICE*, *FactoExtra*, and *FactoMineR* are R packages that implement the multiple imputation method, FAMD, and cluster analysis. The 4M consensus clustering was implemented using R function built by the author. P values < 0.05 were considered statistically significant.

Table 6.2: Proposed framework for multiple imputation in cluster analysis

Input: Eosinophilic dataset with missing values and multidimensional variables
Step 1. Multiple Imputation
i) Obtain 100 complete datasets by multiple imputation (MICE)
Step 2. Factor analysis for mixed data (FAMD)
i) Determine quantitative and qualitative variable
ii) Apply FAMD for each imputed dataset
iii) Determine the number of components for each imputed dataset
Step 3. Hierarchical clustering
i) Choosing the best number of clusters for each imputed dataset
Step 4. Partitioning Clustering
i) Consider the number of clusters in the previous step for each imputed dataset
ii) Assign patients to each cluster for each imputed dataset.
Step 5. 4M method for Consensus Clustering
i) Combine all ensemble clustering to get a final best clustering
Output: Partition of clustering labels $C = \{C_1, \ldots, C_k\}$
Step 6. Assign patients to the final result of consensus clustering
i) Allocate patients in the original incomplete dataset to calculate final result of consensus clustering
Step 7. Description of clustering
i) Calculate median for the original incomplete dataset
ii) Comparison between cluster (Mann-Whitney and Chi-squared tests)
Output: Descriptive analysis tables for clustering

6.3 Clustering Results

Applying the framework detailed in Table 6.2, provided clustering results as follows. Using FAMD as the second step, the percentage contribution of variables could be calculated for the next clustering. Table 6.4 provides the impact of each variable on clustering for the whole cohort, as well as according to ICS treatment. Spirometric parameters, asthma control, circulating granulocytes, atopic status, and age were the variables that contributed the most. The highest contribution for variables in clustering in the whole cohort was for FEV1 pre (% predicted), FEV1 post (%predicted), ACQ, FEV1/FVC post (%), and FVC post (% predicted). Figure 6.2 illustrates the order of contributions of variables from highest to lowest for the whole cohort.

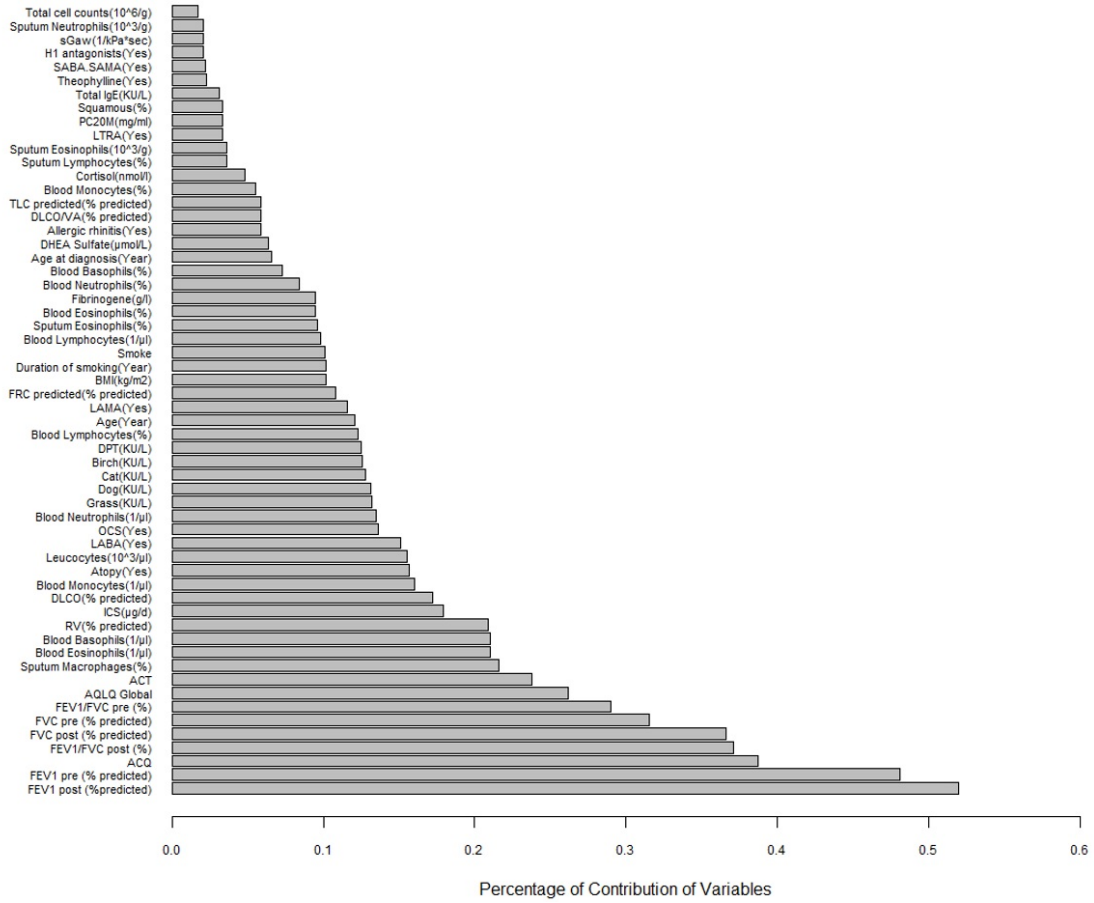

Figure 6.2: The percentage contribution of eosinophilic variables in principal components, in the whole cohort and according to the ICS treatment

After applying multiple imputation for whole cohort, one imputed dataset proposed 32, 33, 37, 39 PCA components, 81 imputed datasets suggested 34 new components, and the others were summarized into 35 components. Next, each imputed dataset was classified based on its own selected components. Out of 100 imputed datasets, four, and five clusters were found in two imputed datasets, 81% were classified into 2 clusters, and the rest had three clusters.

Finally, after completing the framework presented in Table 6.2, two clusters with acceptable validation values for the whole cohort and both subgroups were identified in this chapter. The indices for internal validation of clustering are listed in Table 6.3, as well as the two indices for validation of clustering stability.

Cluster analysis revealed two clusters identified as cluster 1 (n = 276) and cluster 2 (n = 150) (Table 6.5). Cluster 1 included younger patients (50 years), with a high proportion of atopic patients (67%), lower treatment burden (median ICS dose 500 μ g equivalent beclomethasone/d), and preserved lung function (median FEV1 93% predicted) a relatively good asthma control (median ACT and ACQ 18 and 1.3, respectively).

Cluster 2 included older patients (59 years) with a low proportion of atopic (36%), a more frequent smoking history (57%), a higher treatment burden (median ICS dose 2000 g/d equivalent beclomethasone), a more intense systemic and airway eosinophilic inflammation (median circulating eosinophils 379/μ l, median sputum eosinophil count 16%), greater systemic inflammation as reflected by higher fibrinogen levels and circulating neutrophil counts, greater airway obstruction (median FEV1 65% predicted, median FEV1/FVC 69% and median sGaw 0.6 1/KPas.sec) and poorly controlled asthma (median ACT, and ACQ; 11 and 3.1, respectively).

Table 6.3: Clustering validation

	Indices	Whole cohort	Steroid naïve cohort	High dose ICS cohort
Internal measures	Silhouette Coefficient	0.78	0.84	0.76
	Dunn Index	0.59	0.71	0.88
Stability measures	Average Proportion of Non-overlap	0.043	0.038	0.04
	Average Distance between Means	0.006	0.009	0.009

Table 6.4: Percentage of the contribution of variables in clustering, globally and according to the ICS treatment

Variable	**Whole cohort (n = 426)**	**Steroid naïve (n = 114)**	**High dose ICS (n = 239)**
Demographic characteristics			
Age (Year)	0.12	0.06	0.08
Sex (Male)	0.01	0.01	0.005
BMI (kg/m^2)	0.10	0.02	0.005
Smoking	0.10	0.02	0.08
Duration of smoking(Year)	0.10	0.02	0.09
Age at diagnosis (Year)	0.07	0.04	0.06
Duration of asthma(Year)	0.001	0.01	0.005
Atopy (Yes)	0.16	0.03	0.13
Comorbidities			
Nasal polyposis (Yes)	0.001	0.001	0.01
Allergic rhinitis (Yes)	0.06	0.02	0.05
GERD (Yes)	0.001	0.004	0.01
Treatment			
ICS ($\mu g/d$)	0.18		
LABA (Yes)	0.15	0.02	0.001
LTRA (Yes)	0.03	0.002	0.005
LAMA (Yes)	0.12	0.03	0.05
Theophylline (Yes)	0.02	0.001	0.01
H_1 antagonists (Yes)	0.02	0.01	0.05
SABA.SAMA (Yes)	0.02	0.02	0.02
OCS (Yes)	0.14	0.01	0.04
Asthma control, exacerbation, and quality of life			
ACT	0.24	0.2	0.08
ACQ	0.39	0.18	0.16
Number of exacerbation	0.01	0.03	0.009
AQLQ Global	0.26	0.21	0.12
Pulmonary function			
FEV1 pre (% predicted)	0.48	0.14	0.24
FEV1 post (%predicted)	0.52	0.13	0.22
FVC pre (% predicted)	0.31	0.14	0.19
FVC post (% predicted)	0.37	0.14	0.17
FEV1/ FVC pre (%)	0.29	0.15	0.12
FEV1/ FVC post (%)	0.37	0.18	0.14
FRC (% predicted)	0.11	0.7	0.05
DLCO/VA (% predicted)	0.06	0.63	0.08
DLCO (% predicted)	0.17	0.66	0.10
RV (% predicted)	0.21	0.72	0.05
TLC (% predicted)	0.06	0.69	0.04
sGaw (1/kPa*sec)	0.02	0.74	0.08
PC20M(mg/ml)	0.03	0.002	0.11
FeNO (ppb)	0.02	0.003	0.03

Table 6.4 – continued from previous page

Variable	Whole cohort (n = 426)	Steroid naïve (n = 114)	High dose ICS (n = 239)
Sputum cell count			
Weight of sputum (g)	0.01	0.19	0.01
Total cell counts (10^6/g)	0.02	0.004	0.02
Viability (%)	0.001	0.03	0.003
Squamous (%)	0.03	0.003	0.005
Macrophages (%)	0.22	0.006	0.006
Lymphocytes (%)	0.04	0.003	0.003
Neutrophils (%)	0.01	0.005	0.02
Eosinophils (%)	0.1	0.01	0.03
Epithelial cells (%)	0.005	0.004	0.003
Macrophages (10^3/g)	0.02	0.01	0.11
Lymphocytes (10^3/g)	0.002	0.03	0.01
Neutrophils (10^3/g)	0.02	0.02	0.005
Eosinophils (10^3/g)	0.04	0.004	0.06
Epithelial cells (10^3/g)	0.002	0.003	0.01
Blood cell count			
Leucocytes ($10^3/\mu$l)	0.15	0.50	0.32
Neutrophils (%)	0.08	0.62	0.41
Lymphocytes (%)	0.12	0.63	0.54
Monocytes (%)	0.05	0.69	0.45
Eosinophils (%)	0.09	0.67	0.47
Basophils (%)	0.07	0.78	0.55
Neutrophils (1/μl)	0.13	0.64	0.53
Lymphocytes (1/μl)	0.1	0.62	0.43
Monocytes (1/μl)	0.16	0.69	0.51
Eosinophils (1/μl)	0.21	0.64	0.50
Basophils (1/μl)	0.21	0.73	0.52
Serum IgE			
RAST Birch (t3) %>0.35 (KU/L)	0.13	0.05	0.08
RAST Mould (MIX1) %>0.35 (KU/L)	0.01	0.07	0.03
RAST Grass (GX3) %>0.35 (KU/L)	0.13	0.04	0.13
RAST Dog (e5) %>0.35 (KU/L)	0.13	0.03	0.13
RAST Cat (e1) %>0.35 (KU/L)	0.13	0.05	0.11
RAST DPT (d1) %>0.35 (KU/L)	0.12	0.02	0.11
Total IgE(KU/L)	0.03	0.72	0.05
Systematic inflammation			
CRP (mg/l)	0.005	0.75	0.05
Fibrinogen (g/l)	0.09	0.73	0.08
Adrenal function			
Cortisol (nmol/l)	0.05	0.63	0.49
DHEA sulfate (μmol/L)	0.06	0.61	0.39

Table 6.5: Median (IQR) / Percentage (frequency) in each cluster and comparison between clusters.

Variable	Whole cohort (n = 426)			Steroid naïve (n = 114)			High dose ICS (n = 239)		
	Cluster 1 (n=276)	**Cluster 2 (n=150)**	**P-Value**	**Cluster 1 (n=66)**	**Cluster 2 (n=48)**	**P-Value**	**Cluster 1 (n=120)**	**Cluster 2 (n=119)**	**P-Value**
Demographic characteristics									
Age (Year)	50 (34 - 63)	59 (48 - 65)	<0.0001	41 (28 – 51)	61 (53 - 70)	<0.0001	45 (30 – 58)	59 (50 - 66)	<0.0001
Sex (Male)	47% (129)	41%(61)	0.27	58%(38)	40% (19)	0.09	40%(48)	41% (49)	0.96
BMI (kg/m^2)	26 (23 - 29)	27 (23 – 30)	0.17	25 (22 – 27)	27 (23 - 30)	0.009	26 (23 – 30)	27 (24 - 30)	0.85
Smoking Ex-smoker	24% (65)	41%(61)	<0.0001	9%(6)	42% (20)	0.0002	21%(25)	45% (54)	0.0001
Smoking Smoker	13% (35)	16%(27)		21%(14)	17% (8)		8%(10)	18% (21)	
Duration of smoking(Year)	15 (6 – 26)	20 (9 - 42)	<0.0001	0 (0 – 1)	5 (0 – 24)	0.0001	0 (0 – 5)	6 (0 – 22)	0.0001
Age at diagnosis (Year)	26 (7 - 50)	39 (19 - 54)	.02	18 (5 – 31)	59 (48 - 63)	<0.0001	15 (5 – 35)	45(28 – 54)	<0.0001
Duration of asthma(Year)	14 (3 – 29)	18 (4 – 32)	0.12	17 (3 – 26)	0 (0 – 5)	<0.0001	24 (10 – 36)	15(4 –25)	0.004
Atopy (Yes)	67%(184)	36%(54)	<0.0001	89%(59)	21%(10)	<0.0001	84%(101)	27%(32)	<0.0001
Comorbidities									
Nasal polyposis (Yes)	12% (33)	8% (12)	0.28	6% (4)	6% (3)	0.99	13% (16)	11% (13)	0.75
Allergic rhinitis (Yes)	35% (97)	18% (27)	<0.0001	50% (33)	6% (3)	<0.0001	44% (53)	14% (17)	<0.0001
GERD (Yes)	13% (36)	13% (19)	0.99	12% (8)	12% (6)	0.99	12% (15)	14% (17)	0.81
Treatment									
ICS (μg/d)	500 (0 - 1000)	2000 (1000 - 2000)	<0.0001						
LABA (Yes)	55% (152)	83%(125)	<0.0001	1%(1)	4% (2)	0.71	91%(109)	89% (106)	0.65
LTRA (Yes)	20% (55)	35%(52)	0.001	6%(4)	8% (4)	0.92	43%(52)	32% (38)	0.11
LAMA (Yes)	0.4% (1)	17%(25)	<0.0001	0%(0)	4% (2)	0.30	0.8%(1)	18% (21)	<0.0001
Theophylline (Yes)	0% (0)	3%(5)	0.01	0%(0)	0% (0)	0.09	0.8%(1)	2% (3)	0.60
H_1 antagonists (Yes)	24% (67)	13%(19)	0.006	21%(14)	2% (1)	.006	38%(46)	9% (11)	<0.0001
SABA.SAMA (Yes)	62% (170)	73%(109)	0.004	73%(48)	48% (23)	0.04	67%(80)	71% (84)	0.29
OCS (Yes)	4% (11)	25%(38)	<0.0001	1%(1)	6% (3)	0.4	12%(14)	24% (28)	0.02
Asthma control, exacerbation, and quality of life									
ACT	18 (14 – 21)	11 (8 - 16)	<0.0001	18 (15 - 21)	19 (14 – 22)	0.98	15 (11 - 20)	12 (8 – 17)	<0.0001
ACQ	1.3 (0.7 – 2.1)	3.1 (2.4 – 3.7)	<0.0001	1 (0.7 – 2)	1.4 (0.7 – 2)	0.34	2 (1 – 3)	3 (2 – 4)	<0.0001
Number of exacerbation 1	16% (43)	23%(35)	0.06	14%(9)	15% (7)	0.13	24%(29)	16% (19)	0.21
≥ 2	16% (44)	17%(26)		20%(13)	6% (3)		18%(22)	18% (22)	
AQLQ Global	5.2 (4.1 – 6.1)	3.4(2.7 – 4.5)	<0.0001	5(4 – 6)	5 (4 – 6)	0.55	4.6(3.6– 6)	3.3 (2.6 – 4.5)	<0.0001

Table 6.5 – continued from previous page

Variable	Whole cohort (n = 426)			Steroid naïve (n = 114)			High dose ICS (n = 239)		
	Cluster 1 (n=276)	**Cluster 2 (n=150)**	**P-Value**	**Cluster 1 (n=66)**	**Cluster 2 (n=48)**	**P-Value**	**Cluster 1 (n=120)**	**Cluster 2 (n=119)**	**P-Value**
Pulmonary function									
FEV1 pre (% predicted)	93(83 - 103)	65(52 – 75)	<0.0001	98(87 – 105)	88(78 – 98)	0.009	86(73 – 99)	67(53 – 80)	<0.0001
FEV1 post (%predicted)	99(90 - 109)	71(57 – 81)	<0.0001	104(93 – 112)	95(88 – 104)	0.004	92(79 – 102)	73(59 – 85)	<0.0001
FVC pre (% predicted)	102(93 – 111)	84(73 –93)	<0.0001	105(94 –113)	102(91 – 111)	0.12	92(79 – 102)	79(71 – 89)	<0.0001
FVC post (% predicted)	101(91 – 110)	79(69 - 88)	<0.0001	102(95 - 112)	98(88 – 108)	0.03	98(88 – 109)	84(73 – 94)	<0.0001
FEV1/ FVC pre (%)	77(70 - 82)	65(58 – 73)	<0.0001	77(71 – 82)	73(69 – 77)	0.02	74(67 – 82)	68(59 – 75)	<0.0001
FEV1/ FVC post (%)	80(75 - 85)	69(61 – 75)	<0.0001	83(78 – 86)	76(70 – 81)	0.0003	78(70 – 83)	71(61 – 77)	<0.0001
FRC (% predicted)	114(95 - 134)	128(104 – 143)	0.003	104(91 – 124)	123(105 – 139)	0.01	122(98 – 136)	125(104 – 141)	0.33
DLCO/VA (% predicted)	95(84 – 109)	93(82 – 105)	0.16	95(83 – 111)	79(79 – 107)	0.51	92(84 – 107)	95(82 – 106)	0.6
DLCO (% predicted)	84(74 – 95)	69(58 – 79)	<0.0001	850(76 – 96)	79(67 – 89)	0.05	82(73 – 97)	70(57 – 84)	<0.0001
RV (% predicted)	107(87 – 133)	133(110 – 165)	<0.0001	93(80 – 110)	123(94 – 136)	0.006	122(95 – 154)	131(109 – 161)	0.17
TLC (% predicted)	97(89 - 105)	99(90 – 109)	0.44	95(85 – 102)	99(93 - 112)	0.02	99(89 – 109)	98(89 - 105)	0.28
sGaw (1/kPa*sec)	0.8(0.6 – 1)	0.6(0.4 – 0.9)	<0.0001	0.9(0.8 – 1.2)	0.7(0.6 – 1.1)	0.08	0.7(0.5 – 1.1)	0.6(0.4 – 0.9)	0.07
PC20M(mg/ml)	1.8 (0.62 – 8)	1.2(0.3 – 5)	0.31	1.5(0.5 – 6)	2 (0.7 – 6)	0.73	0.9 (0.6 – 5.8)	2.2 (1.1 – 22)	0.03
FeNO (ppb)	37(21 – 61)	35(18- 62)	0.33	46(26- 78)	30(17 – 59)	0.03	47(26- 78)	34(18 – 52)	0.32
Sputum cell count									
Weight of sputum (g)	2.3(1.3 - 4)	2(1.1 – 3.8)	0.29	3(1 – 4)	2 (1 - 4)	0.19	2.3(1.2 – 4.2)	2.2(1.2 - 3.9)	0.98
Total cell counts (10^6/g)	1.5(0.8 – 3)	2.5(1 – 6)	0.009	1 (0.7 – 2)	3.5(1 – 7)	0.0006	1.4(0.6 – 2.9)	2.5(1 – 7)	0.0006
Viability (%)	71(57- 81)	70(51 – 82)	0.25	61(47 – 76)	78(71- 85)	0.0001	71(58 – 79)	70(56- 83)	0.86
Squamous (%)	13(6 – 30)	10(3 – 21)	0.01	17(9 – 39)	10(6 – 20)	0.02	13(5 – 31)	11(3 – 22)	0.03
Macrophages (%)	27 (15 – 40)	12(6 – 25)	<0.0001	33(21 – 46)	18 (10 – 25)	<0.0001	28(15 – 41)	11 (6 – 24)	<0.0001
Lymphocytes (%)	1(0.4 – 2.2)	0.7(0.2 - 1.8)	0.008	2(0.6 – 4)	1(0.6 – 2)	0.12	0.8(0.2 – 1.8)	0.7(0.2 – 1.8)	0.75
Neutrophils (%)	48(31 – 68)	50(28 – 70)	0.81	37(24 – 53)	65(52 – 77)	<0.0001	49(31 – 69)	53(30 – 69)	0.89
Eosinophils (%)	9(5 – 19)	16(6 – 42)	<0.0001	13(7 – 25)	7(5 – 15)	0.01	8(4 – 16)	18(7 – 45)	<0.0001
Epithelial cells (%)	3(1 – 6)	3(1 – 7)	0.63	4(2 – 8)	2(0.7 – 4)	0.0005	3(2 – 8)	3(1 – 7)	0.18
Macrophages (10^3/g)	391(159 – 892)	290(115 – 762)	0.01	396(152 – 864)	455(159 – 1609)	0.42	338(125 – 864)	323(123 – 754)	0.58
Lymphocytes (10^3/g)	17(4 – 47)	15(1 – 51)	0.21	17(4– 67)	385(16 – 105)	0.11	9(2– 32)	17(2 – 53)	0.11
Neutrophils (10^3/g)	654(252 – 1821)	834(293 – 2774)	0.17	439(184 – 820)	2136(731 – 4453)	<0.0001	532(234 – 1490)	954(326 – 3294)	0.01
Eosinophils (10^3/g)	172(59 – 453)	401(121– 1353)	<0.0001	189(58– 396)	283(104 – 1036)	0.04	112(49– 324)	148(492 – 1561)	<0.0001
Epithelial cells (10^3/g)	54(16 – 131)	67(24 – 177)	0.08	44(14 – 133)	39(20 – 142)	0.84	60(15 – 126)	75(20 – 175)	0.22

Table 6.5 – continued from previous page

Variable	Whole cohort (n = 426)			Steroid naïve (n = 114)			High dose ICS (n = 239)		
	Cluster 1 (n=276)	Cluster 2 (n=150)	P-Value	Cluster 1 (n=66)	Cluster 2 (n=48)	P-Value	Cluster 1 (n=120)	Cluster 2 (n=119)	P-Value
Blood cell count									
Leucocytes ($10^3/\mu$l)	7(6 - 8)	9(7 – 11)	<0.0001	7(6 – 8)	7(6 - 9)	0.003	7(6 – 9)	8(7 – 10)	0.003
Neutrophils (%)	53(47 - 60)	56(49 – 63)	0.008	51(46 – 56)	53(48 - 60)	0.19	56(47 – 62)	55(49 - 62)	0.74
Lymphocytes (%)	34(28 – 40)	30(23 – 36)	<0.0001	36(32 – 41)	33(26 – 38)	0.05	32(25 – 38)	30(24 – 36)	0.14
Monocytes (%)	8(6 - 9)	8(6 – 9)	0.71	7(6 – 9)	8(6 - 10)	0.64	8(7 – 9)	7(6 - 9)	0.50
Eosinophils (%)	3.6(2.5 - 5)	4.4(2.6 – 7)	0.005	3(2 – 4)	3(2 - 5)	0.84	4(3 – 6)	4(3 – 7)	0.12
Basophils (%)	0.5(0.3 - 0.7)	0.6(0.4 – 0.8)	0.01	0.6(0.4 – 0.7)	0.6(0.4 - 0.7)	0.52	0.5(0.3 - 0.7)	0.6(0.4 - 0.8)	0.004
Neutrophils (1/μl)	3701(2867 – 4726)	4663(3577 – 6403)	<0.0001	3305(2644 – 3969)	3916(3290 – 4733)	0.004	4029(3157 – 5279)	4444(3529 – 6106)	0.03
Lymphocytes (1/μl)	2341(1896 – 2769)	2381(1940 - 3004)	0.16	2367(1988 - 2754)	2370(1934 – 2920)	0.83	2355(1794 – 2809)	2368(1983 – 2943)	0.28
Monocytes (1/μl)	533(442 – 669)	666(533 – 813)	<0.0001	512(392 – 610)	589(480 – 746)	0.006	566(471 – 695)	625(493 – 801)	0.06
Eosinophils (1/μl)	242(173 – 377)	379(209 – 627)	<0.0001	220(177 – 292)	259(181 – 375)	0.3	275(191 – 421)	371(190 – 613)	0.02
Basophils (1/μl)	32(22 – 50)	49(32 - 67)	<0.0001	32(22 - 47)	39(28 – 60)	0.1	32(23 – 50)	50(32 – 68)	<0.0001
Serum IgE									
RAST Birch (t3) %>0.35 (KU/L)	28% (78)	7%(11)	<0.0001	44%(29)	2% (1)	<0.0001	37%(45)	2% (3)	<0.0001
RAST Mould (MIX1) %>0.35 (KU/L)	14% (40)	15%(22)	0.999	14%(9)	4% (2)	0.15	27%(33)	9% (11)	<0.0001
RAST Grass (GX3) %>0.35 (KU/L)	39% (108)	15%(23)	<0.0001	58%(38)	0% (0)	<0.0001	55%(66)	8% (9)	<0.0001
RAST Dog (e5) %>0.35 (KU/L)	38% (104)	15%(23)	<0.0001	58%(38)	4% (2)	<0.0001	56%(67)	3% (3)	<0.0001
RAST Cat (e1) %>0.35 (KU/L)	38% (104)	14%(21)	<0.0001	56%(37)	6% (3)	<0.0001	52%(63)	5% (6)	<0.0001
RAST DPT (d1) %>0.35 (KU/L)	47% (130)	21%(32)	<0.0001	65%(43)	10% (5)	<0.0001	62%(74)	13% (16)	<0.0001
Total IgE(KU/L)	215 (66 – 528)	154(54 – 364)	0.03	285(102 – 452)	104(27 – 256)	0.0006	383(103 – 983)	137 (48 – 284)	<0.0001
Systematic inflammation									
CRP (mg/l)	1.7(0.8 – 4)	2.5(1.2– 6.4)	0.03	1(0.6– 4)	2(1 – 3)	0.15	1.7(0.8– 4)	2.8(1.2 – 6.6)	0.01
Fibrinogen (g/l)	3.24(2.7 - 3.6)	3.57(3 – 4)	<0.0001	3.05(2.41 – 3.51)	3.32(2.91 - 3.9)	0.01	3.14(2.59 – 3.61)	3.6(3.1 – 4)	<0.0001
Adrenal function									
Cortisol (nmol/l)	199(156 – 252)	230(178 – 300)	0.02	221.6(189 – 301)	243(195 – 293)	0.89	177(141 – 221)	220(176 – 301)	<0.0001
DHEA sulfate (μmol/L)	2.7(1.6 – 6)	2.4(1 – 3)	0.04	5(3 – 8)	2(2 – 4)	0.08	2.7(1.7 – 5.8)	2.2(1.2 – 3.1)	0.08

[2]BMI(Body Mass Index); GERD(Gastroesophageal Reflux Disease); LABA(Long Acting B2 Agonist); LAMA(Long Acting Muscarinic Antagonist); LTRA(Leukotriene Receptor Antagonist); ICS(Inhaled Corticosteroids); OCS(Oral Corticosteroids); CRP(C-Reactive Protein); DHEAS(Dehydroepiandrosterone Sulfate); IgE(Immunoglobulin E); FENO(Fractional Exhaled Nitric Oxide); FEV1(Forced Expiratory Volume in one second); FVC(Forced Vital Capacity); TLC(Total Lung Capacity); RV(Residual Volume); DLCO(Diffusing Capacity for Carbon Monoxide); FRC(Functional residual capacity)

Clustering on ICS naïve patients ($n = 114$) and those receiving high dose ICS treated patients ($n = 239$) also yielded two clusters that mainly differentiate by age, atopic status, and intensity of granulocytic airway inflammation (Table 6.5). The cluster analysis identified two clusters in ICS naïve patients ($n = 114$), which were termed cluster 1 ($n = 66$) and cluster 2 ($n = 48$). In high dose ICS treated patients ($n = 239$), two clusters were also found, with cluster 1 containing 120 patients, while cluster 2 included 119 patients (Table 6.5).

When comparing the two clusters, in the three cohorts, there was a significant difference in age. Patients in Cluster 2 were significantly older than those in Cluster 1, in the whole cohort (59 vs 50 years old), in the steroid naïve cohort (61 vs 41 years old), and in the high dose ICS treated cohort (59 vs 45 years old) (Figure 6.3). The proportion of atopy in cluster 1 was significantly higher than that in cluster 2, in all three cohorts (Figure 6.4).

While the cohort of steroid naïve patients had relatively preserved lung function in both clusters (median FEV1 of 98% and 88% predicted in cluster 1 and 2 respectively, Figure 6.6) the cohort of patients treated with high dose ICS displayed a marked difference in expiratory flow rates between the two clusters (86% predicted and 67% predicted in cluster 1 and 2 respectively). Cluster 2 of patients treated with high dose of ICS had a more frequent smoking history, greater circulating basophil counts ($p < 0.001$), and, surprisingly, a greater level of cortisol as compared to cluster 1, Figure 6.7 ($p < 0.001$) while the proportion of patients receiving maintenance OCS was higher in cluster 2 (24% vs 12%). The number of patients with adrenal insufficiency (morning cortisol < 102 nmol/l) in patients treated with high dose ICS was, however, similar between the two clusters (22% vs 19% in cluster 1 and 2 respectively). As opposed to cortisol, the dehydroepiandrosterone sulfate levels (DHEAS) were higher in cluster 1 ($p = 0.07$).

ACT was significantly higher in cluster 1 in the whole cohort and the high dose ICS treated cohort, but not in the steroid naïve cohort (Figure 6.5). Asthma control and quality of life were similar in the two clusters of steroid naïve patients but patients from cluster 2 in the cohort of patients with high dose ICS had poorer asthma control and quality of life than patients from cluster 1.

When comparing atopic patients between cluster 1 and cluster 2 on the whole cohort there was an increased sensitization in cluster 1 rate to birch and grass pollens

(39% vs 18% and 58% vs 40% in cluster 1 and 2 respectively, $p < 0.01$ for both), cat (57% vs 39% in cluster 1 and 2 respectively, $p < 0.01$) and dog (56% vs 43% in cluster 1 and cluster 2 respectively, $p < 0.05$) in cluster 1 but a higher sensitization rate to molds in cluster 2 (41% vs 22% in cluster 2 and cluster 1 respectively, $p < 0.05$) while total serum IgE did not differ between the two clusters(median (IQR) $292(113-843)$ vs $349(154-709)$ in cluster 1 and cluster 2 respectively, $p > 0.05$).

6.4 Discussion

After extensive clinical characterization of the patients, we found two clusters among eosinophilic asthmatics that clearly differentiate by demographics, level of asthma control, functional and inflammatory features. The severity of airway obstruction, the level of asthma control, the circulating granulocytes counts as well as atopic status and age were the variables that contribute the most to the clustering.

A cluster with highly atopic patients may show substantial airway eosinophilic inflammation and mild disease as evidenced by a good level of asthma control and preserved lung function despite longer disease duration. By contrast, this cluster seems to be equally at risk of exacerbation in the year prior to the visit as compared to the dominantly non-atopic cluster with severely impaired lung function and poor asthma control, called cluster 2. This suggests that in these atopic patients, eosinophilic airway inflammation may make the patients prone to exacerbate but not necessarily to display a decline in expiratory flow rate over the time. This is even more remarkable that the cluster 1 has a longer disease duration which is in line with earlier disease onset. Our finding is in keeping with the recent study from Hastie et al. who showed that pure eosinophilic phenotype predisposes to exacerbation and health care utilization (Hastie et al., 2021) without leading to lung function decline.

Interestingly, the cluster with severely impaired lung function is a cluster which combines intense airway eosinophilic together with intense neutrophilic inflammation. This finding is in line with previous cross-sectional studies which show that mixed granulocytic inflammation is associated with the worst lung function (Hastie et al., 2010; Schleich et al., 2013; Simpson et al., 2006; Graff et al., 2020). In addition, a recent longitudinal prospective study from the SARP (severe asthma research program) found that patients who combine high sputum and neutrophil airway inflammation are those who display lung function decline over time despite treatment

with ICS (Hastie et al., 2021). There are arguments to suggest that neutrophils are acting as a cofactor to eosinophils to allow them to fully contribute to remodeling (Louis and Schleich, 2021).

Cluster 1 with high atopic proportion had lower sputum and blood neutrophil counts. The sputum neutrophil count was particularly low in Cluster 1 of steroid naïve patients. Whether atopic status may protect against neutrophilic inflammation and remodeling remains unclear (Radermecker et al., 2018) but there is evidence showing that histamine, a mediator released upon IgE mediated activation and basophils, may reduce neutrophil chemotaxis (Bury et al., 1992). Furthermore, one study found that raised sputum tryptase levels in asthmatics were mainly found in patients with selective sputum eosinophilic inflammation with low sputum neutrophil count (Bettiol et al., 1999). Our data also suggest that the possible protection of atopic status against disease severity may differ according to the type of sensitization as cluster 1 includes patients with higher sensitization rate towards birch and grass pollens and animal danders whereas cluster 2 displays a remarkable greater sensitization rate to molds. As we measured specific IgE towards a molds mixture we cannot ascertain whether these patients were specifically sensitized to aspergillus. However, aspergillus is the mold to which asthmatics are the most frequently sensitized and previous studies showed that asthmatics sensitized to aspergillus combined high eosinophilic and neutrophilic airway inflammation, which is one the feature of our cluster 2 (Wark et al., 2000).

Cluster 2 displayed worse asthma control and altered lung function despite higher burden of treatment with higher dose of ICS, a greater proportion of patients with LTRA and maintenance OCS. The persistence of high blood and sputum eosinophil counts in this cluster highlights the inability of corticoids to control eosinophilic inflammation in these asthmatics. This phenomenon partly relates to the inability of corticoids to fully suppress the influence of interleukin -5 (Peters et al., 2019) as we know today that severe eosinophilic asthmatics may dramatically respond to anti-IL-5/IL-5R (Bleecker et al., 2016; Pavord et al., 2012). Interestingly, besides eosinophils, blood basophils were also clearly increased in cluster 2 despite heavy treatment with ICS and sometimes OCS.

Surprisingly, levels of morning cortisol were higher in cluster 2 than in cluster 1 whereas the burden of ICS/OCS was greater in patients of cluster 2, which would suggest some kind of resistance to systemic effect of corticoids on the pituitary/adrenal

axis in the patients from cluster 2. Whether this reduced systemic effect may be, somehow, linked to a reduced sensitivity to the anti-inflammatory effect of corticoids in cluster 2 needs to be further investigated. This hypothesis would, however, be supported by a high airway and circulating levels of cells known to be usually eliminated by corticoids such as eosinophils and basophils. Conversely to cortisol, dehydroepiandrosterone (DHEAS) levels, another adrenal hormone, were higher in cluster 1, an observation likely to be linked to the younger age of patients in cluster 1 (Thomas, 1999).

With respect to comorbidities, there was no difference in GERD and nasal polyposis proportion between the two clusters but allergic rhinitis and anti-H1 consumption were much more frequent in cluster 1 that includes a greater proportion of atopic patients.

The current study presents some limitations. First, the retrospective nature of the study does not allow to be confident on the adherence of the patients nor does it allow to be sure about the accurate number of courses of OCS in the year prior to the visit that defines exacerbation rate. Second, the selection of our eosinophilic phenotype was based on a single sputum analysis whereas it is known that some asthmatics may show intermittent eosinophilic airway inflammation (McGrath et al., 2012). Thus, the considered group of eosinophilic asthmatics may not be entirely representative of a whole eosinophilic asthmatic population. Third, this study is monocentric and should be replicated in other centers using sputum in clinical practice.

We conclude that, among eosinophilic asthmatics, there are two clusters which mainly differentiate by their age, atopic status, their level of functional impairment, the magnitude of granulocytic inflammation, and the level of asthma control. The cluster with the lower proportion of atopic patients is clearly the most severe and resistant to corticoids. Whether eosinophils are phenotypically and functionally different among the two clusters warrant further investigation.

6.5 Appendix

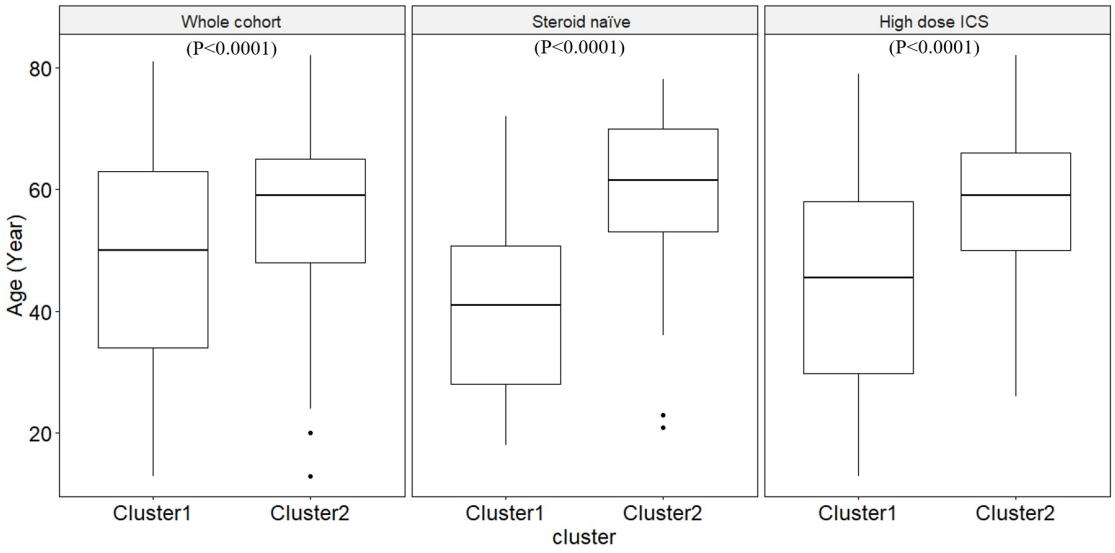

Figure 6.3: Age (Year) for three categories of cohorts in two clusters

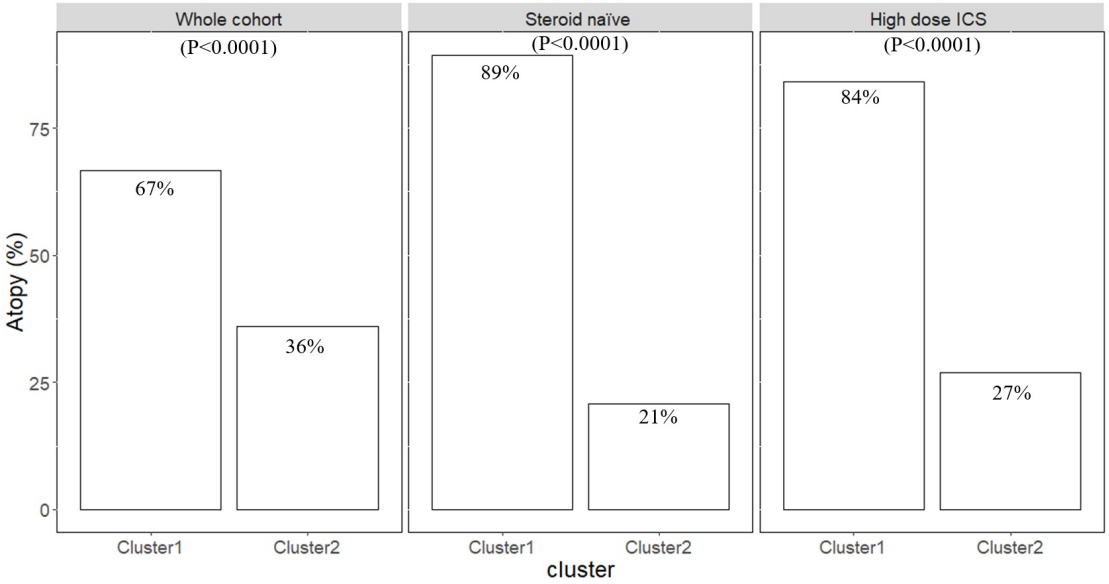

Figure 6.4: Atopy (Yes) for three categories of cohorts in two clusters

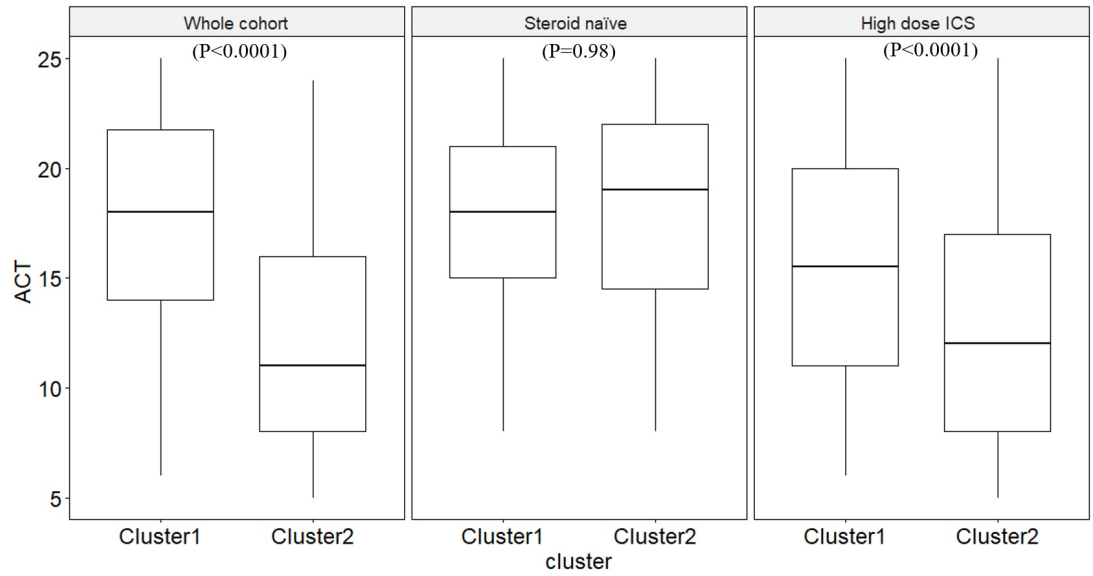

Figure 6.5: ACT for three categories of cohorts in two clusters

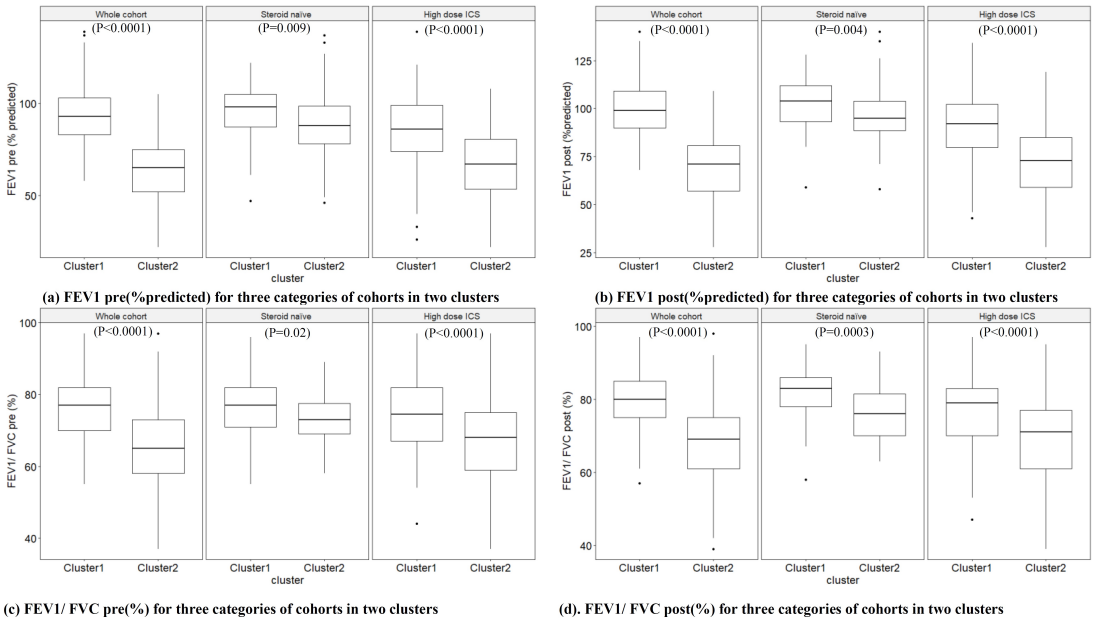

Figure 6.6: Pulmonary function for three categories of cohorts in two clusters

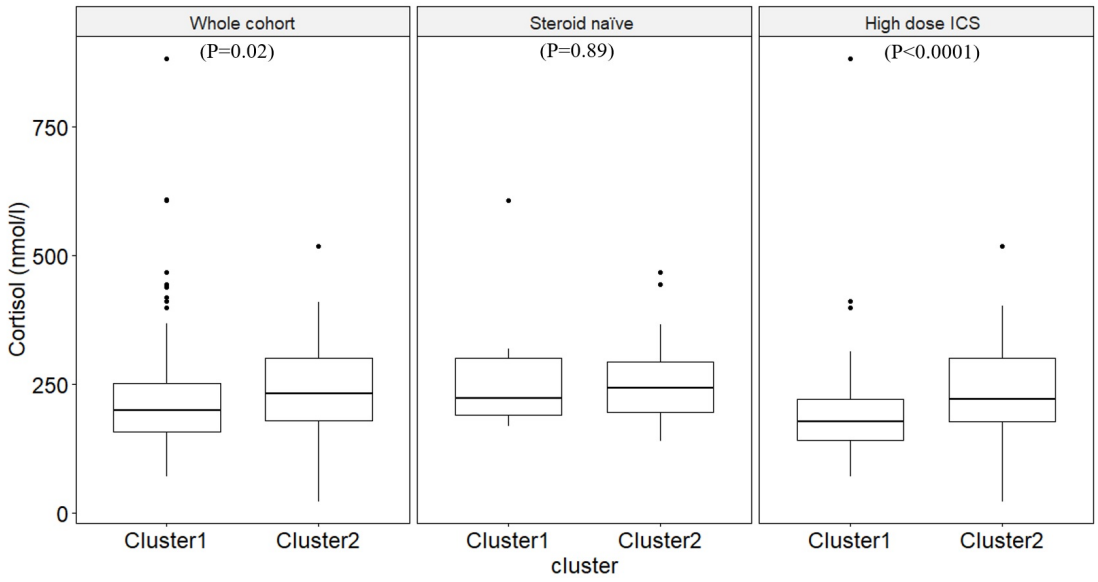

Figure 6.7: Cortisol (nmol/l) for three categories of cohorts in two clusters

CHAPTER 7

Clustering on the non-eosinophilic asthmatic patients

In Chapter 6, we attempted to find more homogenous subgroups among eosinophilic asthmatics. A substantial fraction of asthmatics do not display significant airway eosinophilia. Non-eosinophilic asthma is likely to be composed of different phenotypes. This chapter focuses on performing cluster analyses on a large number (n = 588) of non-eosinophilic (sputum eosinophils <3%) asthmatics. According to the same procedure, patients were split into steroid naïve and high-dose ICS treated. The characteristics of non-eosinophilic asthmatic patients as well as the characteristics of two ICS subsets are presented in Section 7.1, along with a comprehensive explanation of the missing values. The proposed clustering framework is repeated in Section 7.2 for the clustering of the non-eosinophilic asthmatic datasets. The results of clustering for the whole non-eosinophilic asthma cohort, steroid naïve cohort, and high dose ICS treated cohort are presented in Section 7.3. Clinical interpretations and explanations of the cluster analysis results are discussed in Section 7.4.

This Chapter is based on

Nekoee Zahraei, H., Guissard, F., Paulus, V., Henket, M., Donneau, A.F., & Louis, R., Clustering on the non-eosinophilic asthmatic patients. (manuscript in under-review)

7.1 Non-eosinophilic asthmatic dataset

Asthma is chronic airway disease, usually associated with airway inflammation, characterized by the conjunction of respiratory symptoms such as dyspnea, chest tightness and wheezing together with excessive airway caliber fluctuation (GINA 2021). The inflammatory process frequently features an eosinophilic inflammation often combined with a raised IgE production directed against aeroallergens defining the T2 high phenotype. Today, the majority of asthma treatment and prevention strategies are focused on allergic and eosinophilic asthma (O'Byrne et al., 2019; Pavord et al., 2018). However, asthma may be present without airway eosinophilic inflammation (McGrath et al., 2012). While much work has been done on T2 high asthma, the mechanisms leading to a T2 low disease have been much less studied although chronic infection and pollutant exposure are thought to contribute (Douwes, 2002; Esteban-Gorgojo et al., 2018; Fitzpatrick et al., 2020).

Therefore, it is likely that T2 low asthma represents a heterogeneous group of patients. Clustering has become a popular method to identify phenotypes among a large set of asthmatic patients (Bourdin and Chanez, 2013). It is therefore important to study non-eosinophilic asthma further, and to investigate how these patients can be grouped into multiple homogenous clusters.

In this study, three groups of patients were evaluated: i) all non-eosinophilic asthmatics, ii) all patients without ICS, and iii) all patients with ICS >1000 g/d equivalent beclomethasone. In all parts, qualitative variables were presented as count and percentage, while quantitative variables were expressed as median and interquartile range (P25 - P75).

The descriptive table, Table 7.1, also included the number of missing values and their percentages. The percentage of missing values ranged from 0% to 66% for DHEA sulfate (μmol/L) and 94% of patients presented at least one missing value. According to the investigation, the missing values were missing completely at randoms.

In addition, in Figure 7.1, a general perspective of the whole non-eosinophilic asthmatic cohort, and two subgroups were presented. Variables are represented by columns, and patients by rows. Missing data is represented by white cells.

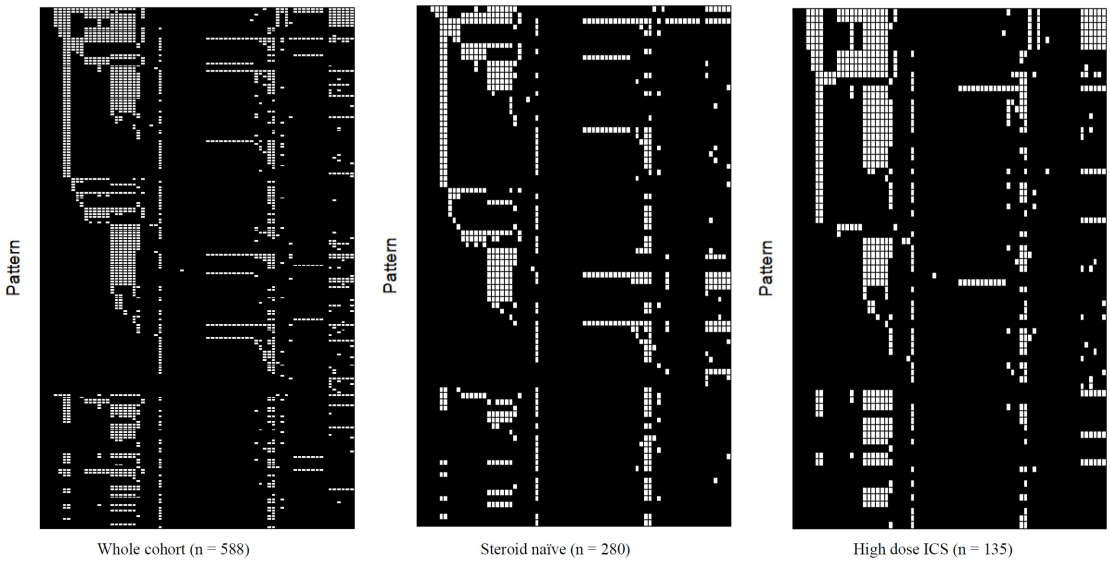

Figure 7.1: A general overview of non-eosinophilic asthmatic cohorts (white cells are missing values)

Table 7.1: Descriptive statistics of characteristics of non-eosinophilic asthmatic cohort

Variable	**Whole cohort(n = 588)**		**Steroid naïve (n = 280)**		**High dose ICS (n = 135)**	
	Median (IQR) %(n)	**Missing Value %(n)**	**Median (IQR) %(n)**	**Missing Value %(n)**	**Median (IQR) %(n)**	**Missing Value %(n)**
Demographic characteristics						
Age (Year)	50(35-61)	1.53% (9)	50(36–61)	0.36% (1)	48(34–58)	4.4% (6)
Sex (Female)	61.90% (364)	4.4% (6)	61% (171)	0.36% (1)	1.53% (9)	58.5% (79)
BMI (kg/m^2)	26(22.85-29.70)	3.57% (21)	24.9(22–28.9)	0.72% (2)	27.1(23.2–20.9)	6.7%(9)
Smoking		2.38% (14)		5.9% (8)		1.43% (4)
Ex-smoker	25% (147)		23.9% (67)		26.7% (36)	
Smoker	23.47% (138)		28.6% (80)		20.7% (28)	
Age at diagnosis (Year)	33(15-52)	30.95% (182)	36(20–53)	32.26% (90)	32(14.2–45.7)	33.3% (45)
Duration of asthma(Year)	5.45(0.72-22)	1.7% (10)	3(0.1–13.6)	32.26% (90)	12(2.6–22.9)	33.3% (45)
Atopy (Yes)	45.58% (268)	1.7% (10)	38.6% (108)	2.5% (7)	49.6% (67)	1.48% (2)
Treatment						
ICS (yes)	50.34% (296)	2.21% (13)				
ICS (μg/d)	0(0-1000)	5.1% (30)				
LABA (Yes)	48.30% (284)	2.21% (13)	1.8% (5)	0% (0)	100% (135)	0% (0)
LTRA (Yes)	20.41% (120)	2.21% (13)	5.7% (16)	0% (0)	44.4% (60)	0% (0)
LAMA (Yes)	4.25% (25)	2.21% (13)	1.8% (5)	0% (0)	9.6% (13)	0% (0)
Theophylline (Yes)	1.02% (6)	2.21% (13)	0% (0)	0% (0)	3% (4)	0% (0)
H_1 antagonists (Yes)	14.12% (83)	2.21% (13)	7.5% (21)	0% (0)	22.2% (30)	0% (0)
SABA.SAMA (Yes)	58.84% (346)	0% (0)	56.1% (157)	0% (0)	73.3% (99)	0% (0)
OCS (Yes)	2.89% (17)	2.21% (13)	1.8% (5)	0% (0)	5.9% (8)	0% (0)
Asthma control, exacerbation, and quality of life						
ACT	16(12-21)	2.04% (12)	18(13–22)	2.5	12(9–16)	1.48% (2)
ACQ	1.7(0.7-2.7)	2.38% (14)	1.3(0.5–2.2)	2.1% (6)	2.6(1.5–3.3)	1.48% (2)
Number of exacerbation		9.35% (55)		11.4% (32)		5.19% (7)
0	64.8% (381)		77.1% (216)		44.4% (60)	
1	15.14% (89)		7.5% (21)		23% (31)	
≥ 2	10.71% (63)		3.9% (11)		27.4% (37)	
AQLQ Global	4.7(3.7-5.7)	3.1% (18)	5.3(4.1–6.1)	2.9% (8)	3.6(2.9–4.8)	1.48% (2)

Table 7.1 – continued from previous page

Variable	Whole cohort(n = 588)		Steroid naïve (n = 280)		High dose ICS (n = 135)	
	Median (IQR) %(n)	Missing Value %(n)	Median (IQR) %(n)	Missing Value %(n)	Median (IQR) %(n)	Missing Value %(n)
Pulmonary function						
FEV1 pre (% predicted)	88(77-100)	5.44% (32)	92(81–101)	3.21% (9)	81(62.5–91)	5.93% (8)
FEV1 post (%predicted)	94(84-105)	5.44% (32)	97(88–106)	3.57% (10)	88(76–100)	5.19% (7)
FVC pre (% predicted)	96(85-107)	5.27% (31)	98(88–108)	3.21% (9)	91(76.7–102)	5.19% (7)
FVC post (% predicted)	99(89-109)	10.03% (59)	99(90–109)	4.64% (13)	94(84–108)	14.8% (20)
FEV1/ FVC pre (%)	76(68-81)	5.27% (31)	77(71–82)	3.21% (9)	73(60.7–79)	5.19% (7)
FEV1/ FVC post (%)	80(73-85)	5.61% (33)	81(75–85)	3.57% (10)	78(70–82.2)	5.19% (7)
FRC (% predicted)	117(96-135)	34.01% (200)	118(96–139)	22.5% (63)	115.5(96.7–143)	49.63% (67)
DLCO/VA (% predicted)	93(79-104)	35.54% (209)	92(76.2–102.7)	23.57% (66)	92(80–108)	50.37% (68)
DLCO (% predicted)	79(67-90)	35.71% (210)	77(65–90)	23.93% (67)	74(61.5–88.5)	50.37% (68)
RV (% predicted)	111(91-136)	33.67% (198)	113(92–136)	22.86% (64)	117(97–148)	48.89% (66)
TLC (% predicted)	98(89-108)	33.33% (196)	98(88–108)	22.5% (63)	97.5(87.2–108.7)	48.15% (65)
sGaw (1/kPa*sec)	0.9(0.6-1.2)	36.22% (213)	0.9(0.7–1.1)	25% (70)	0.8(0.5–1.1)	51.11% (69)
PC20M(mg/ml)	2.45(0.79-6.83)	20.58% (121)	3.06(1.41–8.23)	10% (28)	1.66(0.54–5.02)	44.44% (60)
FeNO (ppb)	18(11-30)	5.27% (31)	18(11.5–26)	3.21% (9)	17.5(9–39.2)	5.19% (7)
Sputum cell count						
Weight of sputum (g)	2.4(1.4-3.9)	0% (0)	2.3(1.5–3.7)	0% (0)	2.3(1.2–3.7)	0% (0)
Total cell counts (10^6/g)	1(0.45-2.6)	0.34% (2)	1.1(0.5–2.6)	0.36% (1)	1.1(0.4–3)	0.74% (1)
Viability (%)	75(58-86)	0.68% (4)	75(59–86)	0.36% (1)	75(59–86)	1.48% (2)
Squamous (%)	18(7-33)	65.14% (383)	18.5(7.0–33.5)	69.3% (194)	22(6.5–36)	62.2% (84)
Macrophages (%)	23.6(11.4-43)	0% (0)	22(10.75–44.12)	0% (0)	24.8(12.1–39.4)	0% (0)
Lymphocytes (%)	0.95(0.2-2)	0% (0)	1(0.2–2.6)	0% (0)	1(0.2–1.8)	0% (0)
Neutrophils (%)	67.6(46.7-83.8)	0% (0)	69.6(43.3–84.7)	0% (0)	65.6(50.6–81.2)	0% (0)
Eosinophils (%)	0.4(0-1)	0% (0)	0.4(0–1)	0% (0)	0.5(0–1.4)	0% (0)
Epithelial cells (%)	2.6(1-6.4)	0.17% (1)	2.3(1–6)	0% (0)	4.0(1.25–7.3)	0.74% (1)
Macrophages (10^3/g)	236.5(95.1-608.3)	0% (0)	240(95.7–638.8)	0% (0)	240(89.7–613.5)	0% (0)
Lymphocytes (10^3/g)	7.62(1.0-28.57)	0% (0)	9.6(1.8–36.4)	0% (0)	7.8(1–30.6)	0% (0)
Neutrophils (10^3/g)	595.4(231.6-1630.8)	0% (0)	621.6(251–173)	0% (0)	595(248–2164)	0% (0)
Eosinophils (10^3/g)	2.6(0-14)	0% (0)	3(0–15.3)	0% (0)	3.6(0–19.2)	0% (0)
Epithelial cells (10^3/g)	29.3(10.5-76.2)	0% (0)	28.9(10–73.6)	0% (0)	36(14.1–93.6)	0% (0)
Blood cell count						
Leucocytes ($10^3/\mu$l)	7(5.9-8.5)	1.36% (8)	7.03(5.9–8.5)	1.43% (4)	7.52(6.09–9.2)	1.48% (2)
Neutrophils (%)	54.95(48.3-61.7)	1.36% (8)	54.7(47.7–60.7)	1.43% (4)	57.5(49.7–65)	1.48% (2)
Lymphocytes (%)	34.1(28.3-40.2)	1.36% (8)	34.8(29.6–40.8)	1.43% (4)	31.2(25.5–39.9)	1.48% (2)
Monocytes (%)	7.6(6.2-9.3)	1.36% (8)	7.7(6.1–9.3)	1.43% (4)	7.3(6.2–9.2)	1.48% (2)
Eosinophils (%)	1.75(1-2.8)	1.36% (8)	1.8(1.1–2.7)	1.43% (4)	1.5(0.7–3.1)	1.48% (2)
Basophils (%)	0.4(0.3-0.6)	1.36% (8)	0.4(0.3–0.6)	1.43% (4)	0.4(0.3–0.6)	1.48% (2)
Neutrophils (1/μl)	3838(2935-4912)	1.36% (8)	3821(2922–4817)	1.43% (4)	4276(3076–5810)	1.48% (2)
Lymphocytes (1/μl)	2345(1896-2927)	1.36% (8)	2390(1911–2900)	1.43% (4)	2343(1960–2980)	1.48% (2)
Monocytes (1/μl)	537(4195-6810)	1.36% (8)	523(400–684)	1.43% (4)	557(442–692)	1.48% (2)
Eosinophils (1/μl)	121(72-191)	1.36% (8)	122(79–191)	1.43% (4)	120(58–200)	1.48% (2)
Basophils (1/μl)	30(20-43)	1.36% (8)	30(19–47)	1.43% (4)	30(20–47)	2.22% (3)
Serum IgE						
RAST Birch (t3) %> 0.35 (KU/L)	16.84% (99)	8.67% (51)	14.3% (40)	5% (14)	15.6% (21)	17.04% (23)
RAST Mould (MIX1) %> 0.35 (KU/L)	7.14% (42)	7.82% (46)	4.6% (13)	4.29% (12)	7.4% (10)	14.07% (19)
RAST Grass (GX3) %> 0.35 (KU/L)	23.13% (136)	9.18% (54)	18.9% (53)	6.07% (17)	24.4% (33)	15.56% (21)
RAST Dog (e5) %> 0.35 (KU/L)	16.5% (97)	6.63% (39)	12.1% (34)	3.57% (10)	17% (23)	14.07% (19)
RAST Cat (e1) %> 0.35 (KU/L)	15.99% (94)	5.95% (35)	12.1% (34)	2.86% (8)	15.6% (21)	12.59% (17)
RAST DPT (d1) %> 0.35 (KU/L)	30.61% (180)	7.99% (47)	27.5% (77)	5% (14)	27.4% (37)	14.07% (19)
Total IgE(KU/L)	69.9(22-197.7)	3.91% (23)	49(18–157)	1.43% (4)	93(36–268)	10.37% (14)
Systematic inflammation						
CRP (mg/l)	2.20(0.93-5.10)	3.91% (23)	2.16(0.92–4.62)	5% (14)	2.41(1.08–5.73)	2.96% (4)
Fibrinogen (g/l)	3.22(2.71-3.72)	56.12% (330)	3.21(2.71–3.7)	3.21% (9)	3.27(2.71–3.79)	2.96% (4)
Adrenal function						
Cortisol (nmol/l)	206.5(149.7-266)	56.12% (330)	232(178–277)	67.86% (190)	168.9(125-235)	46.67% (63)
DHEA sulfate (μmol/L)	3(2-6)	65.65% (386)	4(2–7)	69.3%(194)	3(2–5)	62.96% (85)

[1] BMI(Body Mass Index); LABA(Long Acting B2 Agonist); LAMA(Long Acting Muscarinic Antagonist); LTRA(Leukotriene Receptor Antagonist); ICS(Inhaled Corticosteroids); OCS(Oral Corticosteroids); CRP(C-Reactive Protein); DHEAS(Dehydroepiandrosterone Sulfate); IgE(Immunoglobulin E); FENO(Fractional Exhaled Nitric Oxide); FEV1(Forced Expiratory Volume in one second); FVC(Forced Vital Capacity); TLC(Total Lung Capacity); RV(Residual Volume); DLCO(Diffusing Capacity for Carbon Monoxide); FRC(Functional residual capacity)

According to Table 7.1, the whole cohort (n = 588), patients were mainly females (62%) with a median age of 50, and 46% of patients were atopic. Patients displayed a slight overweight (median body mass index was 26). The median asthma duration was 5 years. There were 51% of patients who had never smoked while 25% were ex-

smoker and 23.5% were currently smokers (Table 7.1). There are 65% of patients who had not experienced an exacerbation the year prior to the visit, 15% have had it once, and 11% have had it more than once. Lung function was preserved with spirometric indices within the normal range in the majority of patients.

As shown, in Table 7.1, 280 steroid naïve non-eosinophilic asthmatic patients were included with a median age of 50 and age at diagnosis of 36 years. This cohort featured a clear dominance of female gender (61%), a majority of patients with a smoking history (52%) and a minority of atopic (39%) patients. The median BMI was 24.9. 77% of patients had not experienced an exacerbation the year prior to the study, 7% had experienced it once, and 4% had experienced it more than once. 135 patients with non-eosinophilic asthma received high-dose ICS. The patients were mostly female (59%) with a median age of 48 and median age asthma duration was 12 years. Half of this cohort (50%) were atopic patients. Patients displayed a mild overweight (median body mass index was 27.1). 44% of patients had not experienced an exacerbation the year prior to the study, 23% had experienced it once, and 27% had experienced it more than once.

7.2 Clustering Framework

The procedure for applying cluster analysis and consensus clustering to consider the uncertainty of missing values in analysis, and reduce the dimension of variables were the same as that performed for the eosinophilic asthmatic cohort and described in detail in Chapter 6. The framework was summarized in Table 7.2.

Table 7.2: Proposed framework for multiple imputation in cluster analysis

Input: Non- eosinophilic dataset with missing values and multidimensional variables
Step 1. Multiple Imputation (Section 3.1.3)
i) Obtain 100 complete datasets by multiple imputation (MICE)
Step 2. Factor analysis for mixed data (FAMD) (Section 3.2.2)
i) Determine quantitative and qualitative variable
ii) Apply FAMD for each imputed dataset
iii) Determine the number of components for each imputed dataset
Step 3. Hierarchical clustering (Section 2.2.2)
i) Choosing the best number of clusters for each imputed dataset
Step 4. Partitioning Clustering (Section 2.2.1)
i) Consider the number of clusters in the previous step for each imputed dataset
ii) Assign patients to each cluster for each imputed dataset.
Step 5. 4M method for Consensus Clustering (Section 3.3.3)
i) Combine all ensemble clustering to get a final best clustering
Output: Partition of clustering labels $C = \{C_1, \ldots, C_k\}$
Step 6. Assign patients to the final result of consensus clustering
i) Allocate patients in the original incomplete dataset to calculate final result of consensus clustering
Step 7. Description of clustering
i) Calculate median for the original incomplete dataset
ii) Comparison between cluster (Mann-Whitney and Chi-squared tests)
Output: Descriptive analysis tables for clustering

7.3 Clustering Results

Based on the framework detailed in Table 7.2, the clustering result is as follows. The percentage contribution of variables could be calculated for the next clustering using FAMD as the second step. For the whole cohort and ICS treated separately, Table 7.4 gives an overview of the impact of each variable on clustering.

In addition, the relative contribution of the variables to the clustering for the whole cohort is detailed in Figure 7.2. With more than 90% contribution, lymphocytes (10^3/g), sputum epithelial cells (%), macrophages (10^3/g), eosinophils (10^3/g), ACT, ACQ, SABA.SAMA, OCS, age, BMI, number of exacerbations, sex, and AQLQ Global were the most influential variables (Figure 7.2).

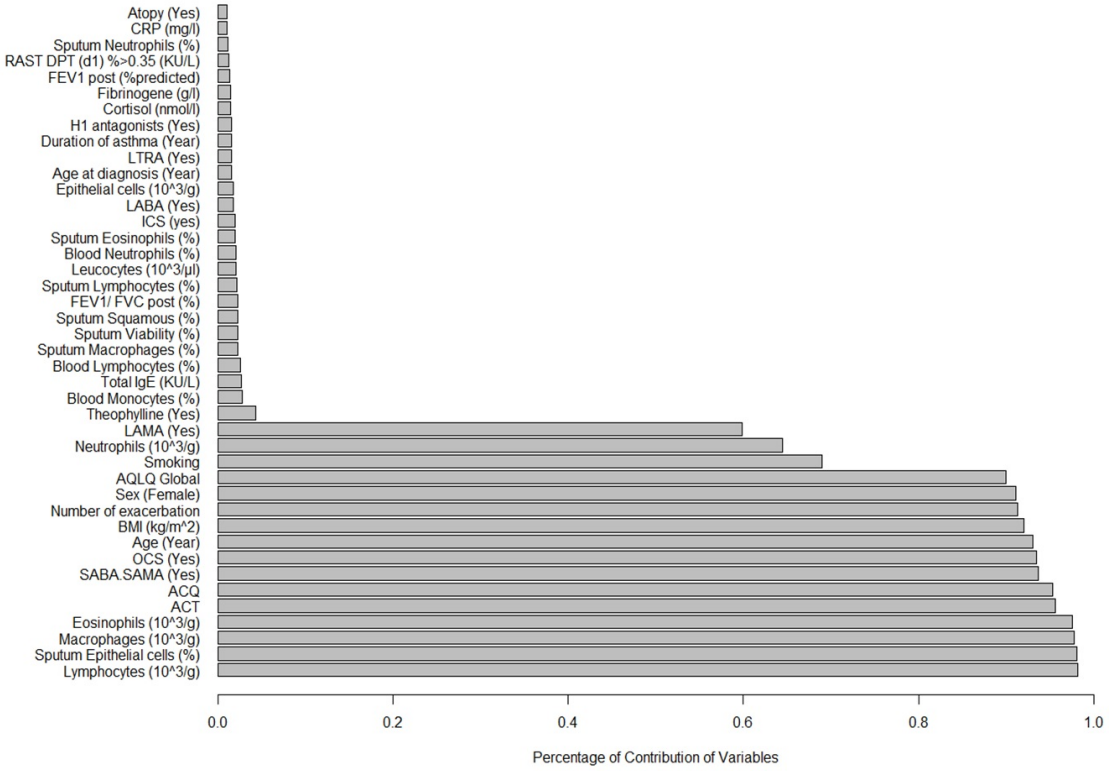

Figure 7.2: Percentage contribution of variables to principal components in the whole cohort

After applying multiple imputation for whole cohort, one imputed dataset proposed 22, 23, 27, and 35 PCA components, 34 imputed datasets suggested 32 new components, and the others were summarized into 33 components. Next, each imputed dataset was classified based on its own selected components. Out of 100 imputed datasets, two clusters were found in 96% imputed datasets, and the rest, 4%, had three clusters.

Finally, as a result of completing the framework presented in Table 7.2, for the whole cohort and both subgroups, the two clusters have accepted validation values in this chapter. In Table 7.3 is listed the indices for internal validation of clustering, as well as the two indices for stability validation.

Table 7.3: Clustering validation

	Indices	Whole cohort	Steroid naïve cohort	High dose ICS cohort
Internal measures	Silhouette Coefficient	0.67	0.88	0.72
	Dunn Index	0.61	0.82	0.85
Stability measures	Average Proportion of Non-overlap	0.028	0.044	0.03
	Average Distance between Means	0.03	0.011	0.01

Table 7.4: percentage contribution of non-eosinophilic variables in principal components, in the whole cohort and according to the ICS treatment

Variable	**Whole cohort (n = 588)**	**Steroid naïve (n = 280)**	**High dose ICS (n = 135)**
Demographic characteristics			
Age (Year)	0.93	0.19	0.20
Sex (Female)	0.91	0.02	0.01
BMI (kg/m^2)	0.92	0.20	0.15
Smoking	0.69	0.04	0.0001
Age at diagnosis (Year)	0.02	0.13	0.13
Duration of asthma(Year)	0.02	0.11	0.12
Atopy (Yes)	0.01	0.16	0.15
Treatment			
ICS ($\mu g/d$)	0.02	0.19	0.30
LABA (Yes)	0.02	0.19	0.31
LTRA (Yes)	0.02	0.19	0.34
LAMA (Yes)	0.60	0.17	0.28
Theophylline (Yes)	0.04	0.03	0.0001
H_1 antagonists (Yes)	0.02	0.05	0.07
SABA.SAMA (Yes)	0.94	0.01	0.0001
OCS (Yes)	0.93	0.12	0.13
Asthma control, exacerbation, and quality of life			
ACT	0.96	0.20	0.1
ACQ	0.95	0.02	0.01
Number of exacerbation	0.91	0.01	0.0001
AQLQ Global	0.90	0.02	0.04
Pulmonary function			
FEV1 pre (% predicted)	0.001	0.01	0.0001
FEV1 post (%predicted)	0.014	0.06	0.08
FVC pre (% predicted)	0.001	0.0001	0.01
FVC post (% predicted)	0.000	0.07	0.08
FEV1/ FVC pre (%)	0.003	0.01	0.01
FEV1/ FVC post (%)	0.022	0.02	0.0001
FRC (% predicted)	0.001	0.12	0.17
DLCO/VA (% predicted)	0.0001	0.01	0.0001
DLCO (% predicted)	0.001	0.09	0.14
RV (% predicted)	0.0001	0.0001	0.0001
TLC (% predicted)	0.001	0.0001	0.01
sGaw (1/kPa*sec)	0.0001	0.01	0.0001
PC20M(mg/ml)	0.0001	0.07	0.06
FeNO (ppb)	0.0001	0.09	0.11

Table 7.4 – continued from previous page

Variable	Whole cohort (n = 588)	Steroid naïve (n = 280)	High dose ICS (n = 135)
Sputum cell count			
Weight of sputum (g)	0.0001	0.01	0.0001
Total cell counts (10^6/g)	0.0001	0.06	0.06
Viability (%)	0.02	0.07	0.08
Squamous (%)	0.02	0.09	0.08
Macrophages (%)	0.02	0.09	0.08
Lymphocytes (%)	0.02	0.01	0.01
Neutrophils (%)	0.01	0.001	0.0001
Eosinophils (%)	0.02	0.001	0.0001
Epithelial cells (%)	0.98	0.10	0.06
Macrophages (10^3/g)	0.98	0.70	0.06
Lymphocytes (10^3/g)	0.98	0.70	0.06
Neutrophils (10^3/g)	0.64	0.69	0.06
Eosinophils (10^3/g)	0.98	0.71	0.02
Epithelial cells (10^3/g)	0.02	0.01	0.0001
Blood cell count			
Leucocytes ($10^3/\mu$l)	0.02	0.07	0.0001
Neutrophils (%)	0.02	0.12	0.23
Lymphocytes (%)	0.03	0.02	0.011
Monocytes (%)	0.03	0.03	0.08
Eosinophils (%)	0.01	0.05	0.04
Basophils (%)	0001	0.03	0.02
Neutrophils (1/μl)	0005	0.01	0.001
Lymphocytes (1/μl)	0.01	0.0001	0.001
Monocytes (1/μl)	0.003	0.28	0.23
Eosinophils (1/μl)	0.001	0.0001	0.001
Basophils (1/μl)	0.004	0.0001	0.0001
Serum IgE			
RAST Birch (t3) %> 0.35 (KU/L)	0.0001	0.01	0.001
RAST Mould (MIX1) %>0.35 (KU/L)	0.001	0.08	0.09
RAST Grass (GX3) %>0.35 (KU/L)	0.002	0.04	0.05
RAST Dog (e5) %>0.35 (KU/L)	0.0001	0.13	0.13
RAST Cat (e1) %>0.35 (KU/L)	0.002	0.06	0.09
RAST DPT (d1) %>0.35 (KU/L)	0.01	0.25	0.23
Total IgE(KU/L)	0.03	0.15	0.08
Systematic inflammation			
CRP (mg/l)	0.01	0.32	0.25
Fibrinogen (g/l)	0.01	0.23	0.14
Adrenal function			
Cortisol (nmol/l)	0.01	0.22	0.17
DHEA sulfate (μmol/L)	0.01	0.33	0.28

Eventually, two subgroups were identified from the cluster analysis: cluster 1 that comprised the large majority of patients ($n = 417$) and cluster 2 ($n = 171$) (Table 7.5).

Patients in cluster 1 were 53 years old, (Figure 7.3), had a low proportion of atopic status (24%), a low level of treatment (55% without ICS) and preserved lung function

(median FEV1 88% predicted) but uncontrolled asthma for the majority of them (median ACT and ACQ 16 and 1.7 respectively, Figure 7.4). Twenty four percent of them reported at least one exacerbation in the previous year.

More than half the patients in cluster 1 had a smoking history with 28.5% ex-smokers and 25% current smokers (Figure 7.3). Patients from this cluster displayed dominant airway and systemic neutrophilic inflammation with median (IQR) sputum and blood neutrophil reaching 70% and 56% respectively. Despite the frequent smoking history in this cluster, spirometric values were preserved in almost 75% (upper three quartiles) of the patients with median FEV1% predicted and FEV1/FVC% reaching 88% and 75% respectively.

Cluster 2 included younger patients (median age 39 years) being almost exclusively atopic (99%) and reporting infrequent smoking history (36%) (Figure 7.3). They had a higher treatment burden with 64% of patients receiving maintenance ICS (median ICS dose 1000 μ g/d equivalent beclomethasone) and had better-controlled asthma (Median ACT and ACQ 18 and 1.3 respectively Figure 7.4) than in cluster 1. The patients had a good lung function with a median FEV1 of 92% predicted and displayed paucigranulocytic asthma for the majority of them.

While spirometric indices were similar between the two clusters, bronchial hyperresponsiveness was more marked in cluster 2 than cluster 1 (Median PC20 M 1.6 vs 2.7, p=0.001). Diffusing capacity and transfer coefficient were slightly altered in cluster 1 and significantly lower than in cluster 2.

Total serum IgE levels were higher in cluster 2 than in cluster 1 (median (IQR) 170 (51 -427) vs 35 (13-103) p< 0.0001), Figure 7.10). Similarly, sensitization rate towards common aeroallergens were much higher in cluster 2 than in cluster 1 (Table 7.5). FeNO levels were significantly higher in cluster 2 than in cluster 1 (25 ppb vs 15 ppb) though median value in cluster 2 remained within the normal range. In contrast to FeNO, fibrinogen levels were significantly higher in cluster 1 than in cluster 2.

Clustering on ICS naïve patients (n = 280) and those treated with high dose ICS treated patients (n = 135) also revealed two clusters that differ mainly by age, disease onset and atopic status (Table 7.5). While in the cohort of steroid naïve patients there was a significant difference in sputum neutrophils (% & 10^3/g) between the two clusters (median neutrophils of 75% in cluster 1 vs 54% in cluster 2).

Table 7.5: Median (IQR) / Percentage (frequency) in each cluster and comparison between clusters.

Variable	Whole cohort(n = 588)			Steroid naïve (n = 280)			High dose ICS (n = 135)		
	Cluster 1 (n=417)	Cluster 2 (n=171)	P-Value	Cluster 1 (n=203)	Cluster 2 (n=77)	P-Value	Cluster 1 (n=89)	Cluster 2 (n=46)	P-Value
Demographic characteristics									
Age (Year)	53(40-62)	39(28-51)	<0.0001	53(39-62)	39.5(28.7-52)	<0.0001	51(40-61)	37(28-49)	<0.0001
Sex (Female)	67.9%(283)	47.4%(81)	<0.0001	67%(136)	45.4%(35)	0.002	64%(57)	47.8%(22)	0.311
BMI (kg/m^2)	26.3(22.8-29.8)	25.4(22.9-29.7)	0.398	25.2(22.2-29)	24.7(21.7-28.7)	0.337	28.1(23.2-31.6)	25.7(23.2-28.3)	0.075
Smoking			<0.001			0.485			0.022
Ex-smoker	28.5%(119)	16.4%(28)		25.6%(52)	19.48%(15)		31.5%(28)	17.4%(8)	
Smoker	25.2%(105)	19.3%(33)		29.1%(59)	28.57%(22)		25.8%(23)	10.9%(5)	
Age at diagnosis (Year)	41(25-56)	17(5-32)	<0.0001	42(31-55)	20(5-35)	<0.0001	37.5(22.2-49)	12.5(5-32)	0.001
Duration of asthma(Year)	2.9(0.3-13.6)	15.5(4.4-30.2)	<0.0001	1.2(0-5.5)	12.5(3-25.9)	<0.0001	8.1(1.9-22.2)	15.7(11.9-25)	0.035
Atopy (Yes)	23.7%(99)	98.8%(169)	<0.0001	19.7%(40)	88.3%(68)	<0.0001	29.2%(26)	89.1%(41)	<0.0001
Treatment									
ICS (yes)	44.6%(186)	64.3%(110)	<0.0001						
ICS ($\mu g/d$)	0(0-1000)	800(0-1600)	<0.0001						
LABA (Yes)	41.9%(175)	63.7%(109)	<0.0001	1.48%(3)	1.3%(1)	0.999	100%(89)	100%(46)	0.0002
LTRA (Yes)	15.1%(63)	33.3%(57)	<0.0001	5.4%(11)	6.5%(5)	0.93	36%(32)	60.9%(28)	0.009
LAMA (Yes)	4.6%(19)	3.5%(6)	0.732	2%(4)	1.3%(1)	0.999	11.2%(10)	6.5%(3)	0.567
Theophylline (Yes)	0.7%(3)	1.7%(3)	<0.0001	0%(0)	0%(0)	<0.0001	2.2%(2)	4.3%(2)	0.883
H_1 antagonists (Yes)	6.7%(28)	32.16%(55)	<0.0001	6.4%(13)	10.4%(8)	0.364	11.2%(10)	43.5%(20)	<0.0001
SABA.SAMA (Yes)	57.8%(241)	61.4%(105)	0.474	54.2%(110)	61%(47)	0.369	78.6%(70)	63%(29)	0.082
OCS (Yes)	3.1%(13)	2.3%(4)	0.812	2%(4)	1.3%(1)	0.999	6.7%(6)	4%(2)	0.862
Asthma control, exacerbation, and quality of life									
ACT	16(12-20)	18(13-22)	0.003	17(13-21)	19(15-23)	0.031	11.5(9-15)	13(10-19)	0.034
ACQ	1.7(0.7-2.7)	1.3(0.7-2.3)	0.018	1.5(0.5-2.3)	1.17(0.54-2)	0.197	2.8(1.7-3.5)	2.2(1-3)	0.028
Number of exacerbation 0	66.9%(279)	59.6%(102)	0.056	77.8%(158)	74%(57)	0.053	41.6%(37)	50%(23)	0.126
1	14.6%(61)	16.4%(28)		4.9%(10)	14.3%(11)		28.1%(25)	13%(6)	
≥ 2	8.9%(37)	15.2%(26)		3.9%(8)	3.9%(3)		24.7%(22)	32.6%(15)	
AQLQ Global	4.5(3.5-5.7)	5.3(4-6.1)	0.001	5.1(4-6)	5.3(4.5-6.1)	0.137	3.4(2.8-4.1)	4.2(3.3-5.7)	0.004

Table 7.5 – continued from previous page

Variable	Whole cohort(n = 135)			Steroid naïve (n = 280)			High dose ICS (n = 239)		
	Cluster 1 (n=417)	**Cluster 2 (n=171)**	**P-Value**	**Cluster 1 (n=203)**	**Cluster 2 (n=77)**	**P-Value**	**Cluster 1 (n=89)**	**Cluster 2 (n=46)**	**P-Value**
Pulmonary function									
FEV1 pre (% predicted)	88(77-100)	87(79-100)	0.78	91(80-101)	92.5(81.7-101.2)	0.771	82(57-91)	81(67-90)	0.6
FEV1 post (%predicted)	93(82-105)	95(86-106)	0.266	96(87-106)	97(90-106)	0.295	87(74-100)	91(78-100)	0.266
FVC pre (% predicted)	96(84-106)	97(87-109)	0.108	98(87-109)	101(90-108)	0.523	89(74-100)	92(82-106)	0.25
FVC post (% predicted)	98(88-108.7)	101(91.5-111.5)	0.034	98(89-109)	103(93-109)	0.318	94(85-107)	97(84-111)	0.415
FEV1/ FVC pre (%)	75(69-81)	76(68-81)	0.751	77(72-82)	77(70-83)	0.662	72(62-79)	73(58-79)	0.798
FEV1/ FVC post (%)	80(73-85)	80(74.7-85.2)	0.536	81(75-85)	81(76-86)	0.402	77(70-81)	80(71-87)	0.063
FRC (% predicted)	118(95-136)	114(99-130)	0.746	119(96-139)	114(94-133)	0.469	117(98-144)	112(87-134)	0.335
DLCO/VA (% predicted)	91(77-102)	98(88-109.7)	<0.0001	88(74-101)	97(84-106)	0.004	89(78-100)	110(98-122)	0.0003
DLCO (% predicted)	75(63-87)	87(77-95)	<0.0001	75(62-88)	85(75-94)	0.001	67(58-83)	88(77-102)	0.0005
RV (% predicted)	113(93-138)	103(88-130)	0.02	114(93-138)	105(89-131)	0.085	126(100-152)	103(82-120)	0.081
TLC (% predicted)	99(88-108)	96(89.2-106)	0.44	100(87-109)	97(90-104)	0.457	98(88-109)	92(81-108)	0.383
sGaw (1/kPa*sec)	0.9(0.6-1.2)	0.8(0.6-1.1)	0.89	0.9(0.7-1.1)	0.85(0.62-1.07)	0.389	0.67(0.47-1)	0.97(0.48-1.39)	0.339
PC20M(mg/ml)	2.7(0.9-7.7)	1.6(0.6-3.7)	0.001	3.2(1.5-8.3)	2.9(1.2-7.2)	0.490	1.3(0.5-3.8)	2.72(0.53-8.21)	0.306
FeNO (ppb)	16(10-25)	25(14-41)	<0.0001	15(10-22)	25.5(13-41)	<0.0001	13(8-33)	35(16-45)	0.0005
Sputum cell count									
Weight of sputum (g)	2.5(1.4-3.9)	2.2(1.16-3.75)	0.074	2.3(1.5-3.6)	2.3(1.5-3.7)	0.738	2.3(1.3-3.6)	2.6(1.1-4.2)	0.898
Total cell counts (10^6/g)	1.1(0.5-2.6)	0.9(0.42.3)	0.116	1.2(0.5-2.8)	0.9(0.48-2)	0.124	1.5(0.4-3.2)	0.8(0.4-2.1)	0.097
Viability (%)	76(62-86)	69.5(53.7-83)	0.003	78(60-87)	68(56-80)	0.010	75(55-85)	75(60-88)	0.506
Squamous (%)	16(6-29)	25(7-39.25)	0.066	18(7.2-34)	20(6.5-29)	0.907	21(6.25-35.7)	22(7-38)	0.749
Macrophages (%)	21.2(10.6-38.4)	32.4(14.3-50.9)	<0.0001	18.6(9.9-36.1)	37.8(20.6-55)	<0.0001	23.8(11.8-38.4)	26(15.5-45.2)	0.413
Lymphocytes (%)	0.8(0.2-2)	1(0.2-2.2)	0.076	0.9(0.2-2)	1.4(0.5-3)	0.016	1(0.2-1.8)	1.2(0.22-1.8)	0.321
Neutrophils (%)	70(51-84.8)	59.2(39.6-79.8)	0.04	75(53.8-86)	54(35-72)	<0.0001	65(51.6-81)	66.6(43.7-79.7)	0.344
Eosinophils (%)	0.4(0-1)	0.4(0-1.2)	0.165	0.4(0-1)	0.4(0-1.2)	0.316	0.5(0-1.4)	0.4(0-1.4)	0.507
Epithelial cells (%)	2.5(1-6.8)	3.2(1.4-5.7)	0.338	2(1-4.5)	3.6(1.4-6.4)	0.033	3.7(1-7.45)	4(1.9-6.6)	0.559
Macrophages (10^3/g)	215(95.2-560)	261.6(96.9-731.6)	0.312	199(89-556)	410(142-754)	0.032	243(102-616)	156(86.8-567.6)	0.258
Lymphocytes (10^3/g)	7.5(1-28.5)	8.4(1.5-29)	0.539	9(1.3-33)	11(2-42)	0.275	7.2(0.6-33.6)	8.9(1.8-24.9)	0.756
Neutrophils (10^3/g)	698(254-1817)	413(162-1302)	0.007	740(268-1954)	353(181-1020)	0.002	940(258-2301)	407(229-1060)	0.066
Eosinophils (10^3/g)	2(0-14)	3.6(0-14.3)	0.281	2.6(0-16.6)	3.6(0-12)	0.727	4.0(0-20.4)	1.28(0-17.5)	0.361
Epithelial cells (10^3/g)	29(10-80)	27.2(12.3-73.4)	0.905	28.6(9.1-84)	30(11-66)	0.732	37.8(18-96)	34.3(12.9-75)	0.508

Table 7.5 – continued from previous page

Variable	Whole cohort(n = 135)			Steroid naïve (n = 280)			High dose ICS (n = 239)		
	Cluster 1 (n=417)	Cluster 2 (n=171)	P-Value	Cluster 1 (n=203)	Cluster 2 (n=77)	P-Value	Cluster 1 (n=89)	Cluster 2 (n=46)	P-Value
Blood cell count									
Leucocytes ($10^3/\mu$l)	7.2(6-8.7)	6.6(5.6-7.9)	0.004	7.2(6.1-8.6)	6.6(5.5-7.9)	0.046	7.9(6.9-9.9)	6.31(5.42-7.98)	0.0001
Neutrophils (%)	55.5(49.1-61.8)	53.5(46.5-61.5)	0.059	54.7(48.3-61)	55(47-60)	0.485	59.6(53-66.6)	50.5(44.5-60.6)	<0.0001
Lymphocytes (%)	33.9(28.3-39.9)	34.7(28.5-41.2)	0.255	34.8(28.9-40.6)	34(30-41)	0.707	29.7(23.6-35.2)	39.4(29.8-43.8)	<0.0001
Monocytes (%)	7.6(6.1-9.2)	7.6(6.5-9.6)	0.284	7.7(6.1-9.2)	7.4(5.9-9.8)	0.979	7.2(6.2-8.7)	7.9(6.5-9.4)	0.133
Eosinophils (%)	1.7(1-2.6)	1.9(1-3.4)	0.104	1.7(1.1-2.6)	1.9(1.3-2.8)	0.156	1.5(0.7-2.5)	1.5(0.7-3.1)	0.836
Basophils (%)	0.4(0.3-0.6)	0.4(0.3-0.6)	0.84	0.4(0.3-0.6)	0.4(0.3-0.6)	0.901	0.4(0.3-0.6)	0.4(0.3-0.6)	0.665
Neutrophils ($1/\mu$l)	3943(3065-5039)	3563(2809-4733)	0.006	3931(3046-4840)	3557(2751-4588)	0.053	4696(3718-6392)	3200(2650-14733)	<0.0001
Lymphocytes ($1/\mu$l)	2362(1910-2988)	2294(1879-2670)	0.138	2431(1918-2962)	2276(1883-2668)	0.276	2366(1883-2994)	2320(2102-2852)	0.915
Monocytes ($1/\mu$l)	540(430-689)	530(402-661)	0.331	538(417-704)	489(378-673)	0.192	595(453-694)	530(411-613)	0.072
Eosinophils ($1/\mu$l)	119(74-187)	128(69-223)	0.407	121(78-183)	128(87-220)	0.511	128(70-199)	109(40-217)	0.244
Basophils ($1/\mu$l)	31(20-46)	30(18-42)	0.215	31(19-49)	30(19-42)	0.791	32(23-47)	29(17-48)	0.036
Serum IgE									
RAST Birch (t3) %>0.35 (KU/L)	1.7%(7)	53.8%(92)	<0.0001	2%(4)	46.7%(36)	<0.0001	4.5%(4)	37%(17)	<0.0001
RAST Mould (MIX1) %>0.35 (KU/L)	1.4%(6)	21%(36)	<0.0001	1.5%(3)	13%(10)	<0.0001	2.2%(2)	17.4%(8)	0.0001
RAST Grass (GX3) %>0.35 (KU/L)	4.8%(20)	67.8%(116)	<0.0001	3.9%(8)	58.4%(45)	<0.0001	7.9%(7)	56.5%(26)	<0.0001
RAST Dog (e5) %>0.35 (KU/L)	1.9%(8)	52%(89)	<0.0001	1%(2)	41.6%(32)	<0.0001	6.7%(6)	37%(17)	<0.0001
RAST Cat (e1) %>0.35 (KU/L)	1.7%(7)	50.9%(87)	<0.0001	0.5%(1)	42.9%(33)	<0.0001	5.6%(5)	34.8%(16)	<0.0001
RAST DPT (d1) %>0.35 (KU/L)	13.4%(56)	72.5%(124)	<0.0001	10.8%(22)	71.4%(55)	<0.0001	11.2%(10)	58.7%(27)	<0.0001
Total IgE(KU/L)	43(17-116)	246(87.5-571)	<0.0001	35(13-103)	170(51-427)	<0.0001	52(22-148)	301(136-816)	<0.0001
Systematic inflammation									
CRP (mg/l)	2.4(1-4.9)	1.8(0.9-5.4)	0.245	2.4(1-4.9)	1.29(0.76-3.10)	0.012	3.2(1.3-7.2)	1.64(0.82-2.58)	0.001
Fibrinogen (g/l)	3.3(2.8-3.8)	2.9(2.5-3.4)	<0.0001	3.36(2.92-3.78)	2.77(2.41-3.2)	<0.0001	3.4(2.8-4.1)	2.9(2.53-3.44)	0.004
Adrenal function									
Cortisol (nmol/l)	218(150-266)	177(150-265)	0.341	239(186-274)	216(162-316)	0.921	159.1(103.4-235)	172.8(152.2218.9)	0.499
DHEA sulfate (μmol/L)	3(2-5.7)	4(2-7)	0.078	4(2-7)	4(3-6)	0.644	2.14(1.79-4.44)	3(2.63-6)	.091

[2]BMI(Body Mass Index); LABA(Long Acting B2 Agonist); LAMA(Long Acting Muscarinic Antagonist); LTRA(Leukotriene Receptor Antagonist); ICS(Inhaled Corticosteroids); OCS(Oral Corticosteroids); CRP(C-Reactive Protein); DHEAS(Dehydroepiandrosterone Sulfate); IgE(Immunoglobulin E); FENO(Fractional Exhaled Nitric Oxide); FEV1(Forced Expiratory Volume in one second); FVC(Forced Vital Capacity); TLC(Total Lung Capacity); RV(Residual Volume); DLCO(Diffusing Capacity for Carbon Monoxide); FRC(Functional residual capacity)

It was the difference in blood neutrophils that was observed between the clusters in the cohort of patients treated with high dose ICS (median neutrophils of 60% in cluster 1 vs 50% in cluster 2) (Figure 7.7, Figure 7.8).

Overall asthma control and quality of life were generally significantly worse in cluster 1 than in cluster 2 and this was especially evident in the group treated with dose ICS (Figure 7.4). Among patients treated with high doses ICS, 53% of patients from cluster 1 reported at least one exacerbation vs 46% in cluster 1. Raised FeNO levels in cluster 2 as compared to cluster 1 was especially evident in those patients treated with high dose of ICS (Figure 7.6). As for fibrinogen it was more elevated in cluster 1 in both steroid naïve and those treated with high dose of ICS (Figure 7.9).

In further investigation, cluster 1 in the whole cohort, as well as in both steroid naïve and high dose ICS, was broken down according to the smoking history. Overall, as compared to never smokers, smoking patients were slightly older with lower FEV1 and FEV1/FVC ratio and impaired DLCO. They also displayed greater airway and systemic granulocytic inflammation and total serum IgE but lower FeNO. Patients treated with high dose ICS who had no smoking history were dominantly female (80%) and had a high BMI (median (IQR) 30 (24-33)).

7.4 Discussion

Non-eosinophilic asthmatics represent a large proportion of our asthmatics with 588 patients over a total of 1014 patients with successful sputum induction screened from our database (58%). Our data provide evidence for two distinct clusters among non-eosinophilic asthmatics. The cluster 1, which is the dominant cluster, include patients the majority of whom had a late disease onset together with a smoking history, display greater magnitude of neutrophilic airway and systemic inflammation, yet without satisfying functional criteria for COPD. The cluster 2 is a cluster of younger dominantly male patients, consisting of, almost exclusively, atopic patients with a classical sensitization profile for our geographical area, a better level of asthma control and quality of life and a greater use of ICS/LABA, LTRA and H1 antagonist as compared to cluster 1.

The reality of smoking asthma has been firmly established in epidemiological studies (Thomson, 2004), yet there is always a trend in medical community to consider a smoker with chronic respiratory symptoms as a patient with chronic bronchitis and/or COPD rather than as an asthmatic. Our patients here were carefully diagnosed based either on reversibility to salbutamol or on bronchial hyperresponsiveness to methacholine. Their spirometric values were considered as being in the normal range with FEV1 % predicted above 80% and post FEV1/FVC ratio well above 70% in the large majority of patients. Our finding supports the concept of smoking induced airway disease is not necessarily accompanied by COPD (Thomson, 2017).

As a consequence of greater smoking history patients from this cluster had mildly impaired diffusion capacity and rather low FeNO values. The airway inflammatory profile was highly neutrophilic with median values close to those seen in COPD, a finding in keeping with the demonstrated relationship between pack years and sputum neutrophils (Demarche et al., 2016). Airway dysbiosis may be another cause of raised sputum neutrophilia in cluster 1 (Abdel-Aziz et al., 2021). Likewise, the blood neutrophil counts was also raised in that cluster, which combined to a slightly increased fibrinogen levels, points to a low grade systemic inflammation.

Patients from cluster 1 had a poorer asthma control and asthma quality of life compared to patients from cluster 2. However, patients from cluster 1 did not report a greater exacerbation rate the year prior to the visit than those from cluster 2, with 60% to 67% of the patients denying any course of OCS in the 12 preceding months in both clusters. This finding indicates that poor day to day asthma control may not necessarily results into greater exacerbation rate, especially in smoking patients.

Atopic status is usually associated with an eosinophilic trait. The reason why the cluster 2 with such a high atopic prevalence remains non-eosinophilic could due to several factors. First it may reflect the impact of ICS on airway eosinophilic inflammation. Up to two third of patients in cluster 2 were receiving ICS combined to LABA as maintenance treatment. It is highly likely that some of these patients were actually eosinophilic prior to starting their treatment with ICS, a class of drug known to be able to sharply decrease sputum eosinophils (Jatakanon et al., 1998; Lim et al., 1999). Second, the lack of airway eosinophilia may also reflect low allergen exposure in daily life as it is well recognized that allergen contact in the airways of a sensitized patient drives a long lasting eosinophilic infiltrate through mast cell activation (Fahy et al., 1994). Allergen avoidance has been reported to result in a decrease in sputum

eosinophils (Piacentini et al., 1996). FeNO values in the cluster 2 were largely in the normal range which also could support the absence of significant exposure though it may evenly be the consequence of chronic treatment with ICS.

Overall cluster 2 received more often maintenance treatment for their asthma with ICS/LABA combination (65% vs 45% in cluster 1) and LTRA (33% vs 15% in cluster 1) and a greater proportion of patients were receiving H1 antagonist for treating allergic rhinitis (32% vs 7% in cluster 1), a finding in keeping with the atopic status in this cluster.

Our study points to a significant group of steroid naïve patients who are featuring paucigranulocytic asthma and normal FeNO. For this group of patients the best treatment strategy is still unclear as ICS were not found superior to placebo or LAMA (Lazarus et al., 2019) and more research needs to be performed to find the most cost effective treatment strategy.

As it is currently unpractical to apply the induced sputum technique on a large scale in routine practice, blood eosinophils have been advocated as a valuable, but imperfect, proxy to approach sputum eosinophils (Demarche et al., 2017). It is worth noting that blood eosinophil count was also low in our two clusters with median values below 150/μl and 75% of the patients with blood eosinophils count less than 187/μl in cluster 1 and less than 223/μl in cluster 2. These values are close to what is seen in a healthy population (Hartl et al., 2020).

By focusing on the patients treated with high doses of ICS we selected patients deemed to have severe asthma (Chung et al., 2014). Several national and international registries have shown that most of the severe asthmatics display sign of T2 high inflammation (van Bragt et al., 2020; Denton et al., 2021). A recent study investigating rizankizumab a p19 IL-23 receptor antagonist, in severe asthma has provided similar finding as it turned out that the large majority of recruited patients displayed sputum eosinophils, although no inflammatory inclusion criteria was mandatory in that study (Brightling et al., 2021).

The novelty study has however challenged this view showing poor relationship between clinical severity of the disease and the magnitude of the eosinophilic trait (Reddel et al., 2021). The patients in that study were, however, qualified as being asthmatics without any firm lung function criteria needed to ascertain the diagnosis.

Our study, which has included asthmatics with demonstration of either reversibility or bronchial hyperresponsiveness to methacholine, also points to a group of severe non-eosinophilic asthmatics as the patients receiving high doses of ICS and still showing insufficient asthma control may be considered as severe according to the ERS/ATS criteria (Chung et al., 2014).

Smoking certainly contribute to poor control in some of those patients as this addiction makes ICS less efficient (Thomson, 2017) and high BMI may certainly contribute in those who are non-smoking as shown in our study, which confirms the cluster of non-eosinophilic obese female first described by Haldar et al. (Haldar et al., 2008).

It is worth noting that in those severe patients from cluster 2, FeNO levels was in the high zone (35 ppb) while airway eosinophilia was absent. The high FeNO levels may actually reflect the almost exclusive atopic status found in tis cluster since it has been firmly established that atopy favours high FeNO (Gerday et al., 2022). However, it is unusual to have such a dissociation in severe asthmatics while low FeNO together with high sputum eosinophils may often be observed in smoking asthmatics. Why FeNO remains high despite ICS is unclear but may reflect a true resistance to ICS (Couillard et al., 2022).

The strength of our study is the application of new method of clustering on a large cohort of non-eosinophilic asthmatics precisely characterized in terms of lung function and airway and systemic inflammation. Our study has, however, several limitations. First, as this is a real life study, we are uncertain about the compliance of the patients to the treatment thus limiting our interpretation of disease severity. Second, we lack accurate data on comorbidities that may play a role in altering asthma control and quality of life such psychologic disorders, gastro oesophagal reflux or chronic rhinosinusitis (Tay and Hew, 2018; Freitas et al., 2020). Third there was no longitudinal follow-up that could give us insight on the evolution lung function decline or exacerbation trend in our clusters.

In conclusion we have provided evidence for two major clusters among non-eosinophilic asthmatics, one containing the greater number of patients, being associated with a late disease onset, a significant smoking history in the majority of the patients together with signs of airway and systemic neutrophilic inflammation, and another cluster including a majority of young and dominantly male, and most exclusively,

atopic patients essentially featuring paucigranulocytic asthma.

Even if smoking cessation should be the first goal to achieve, more treatment trials need to be done in the dominant smoking cluster, a group of patient which has been neglected in the past, yet representing a frequent situation and a real challenge for its management in clinical practice. Likewise, it is worth further investigating what can be the optimal management strategy in those steroid naive paucigranulocytic and FeNO low asthmatics.

7.5 Appendix

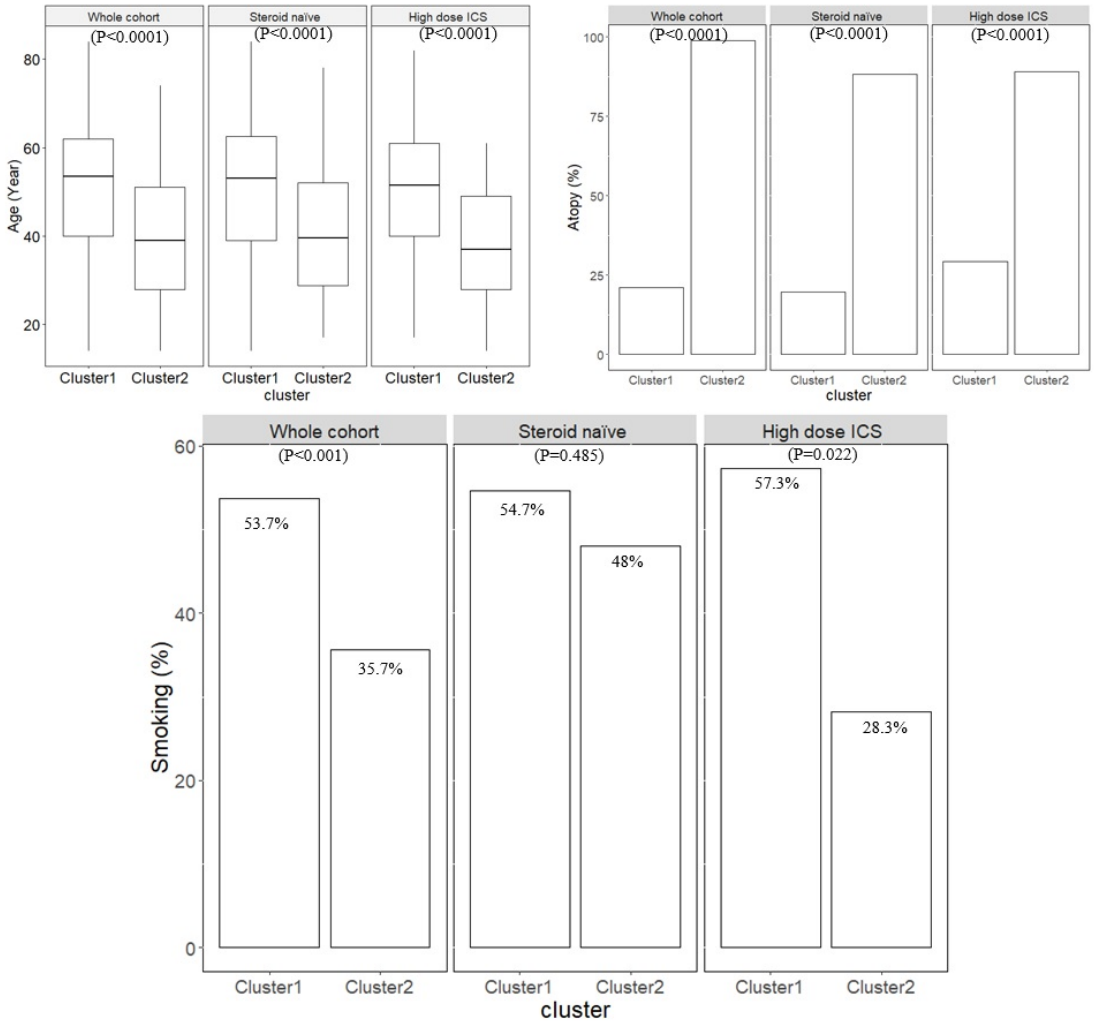

Figure 7.3: Box plot of demographic variables in two clusters of non-eosinophilic asthmatics and their subgroups

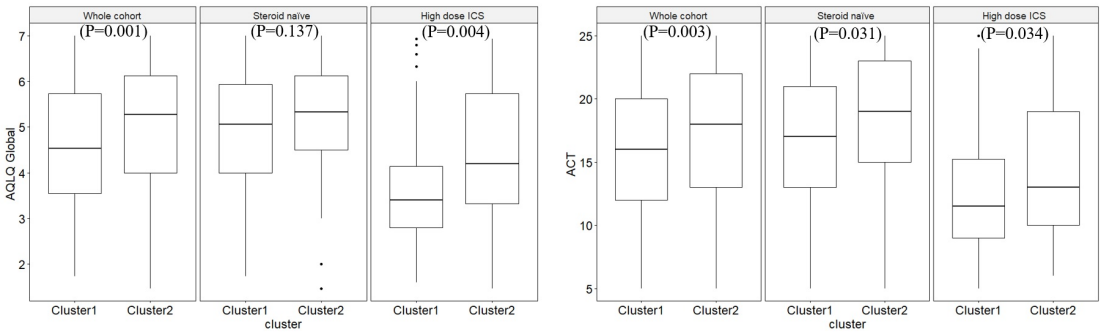

Figure 7.4: ACT and AQLQ for three categories of cohorts in two clusters

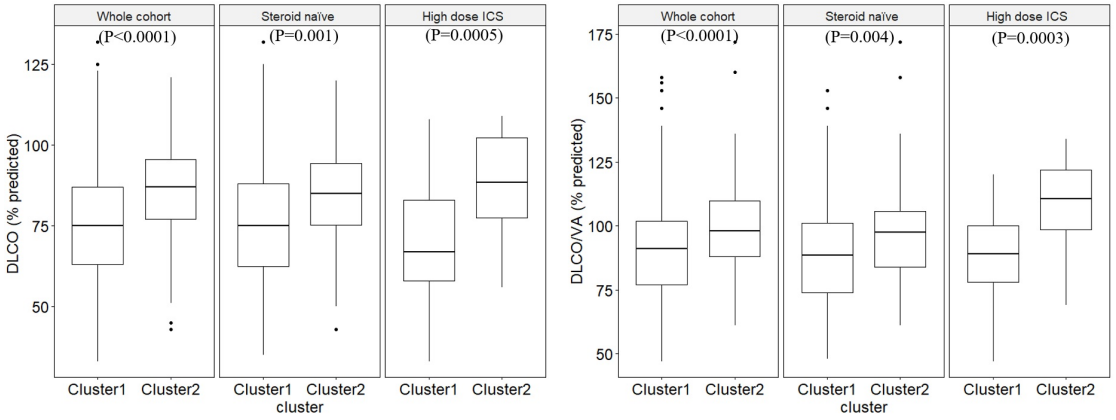

Figure 7.5: DLCO (% predicted) and DLCO/VA (% predicted) for three categories of cohorts in two clusters

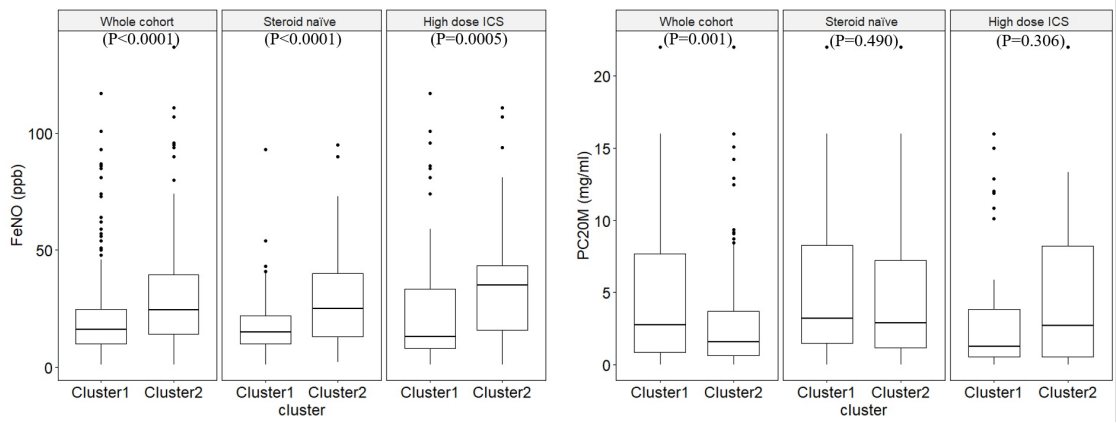

Figure 7.6: FeNo(ppb) and PC20M(mg/ml) for three categories of cohorts in two clusters

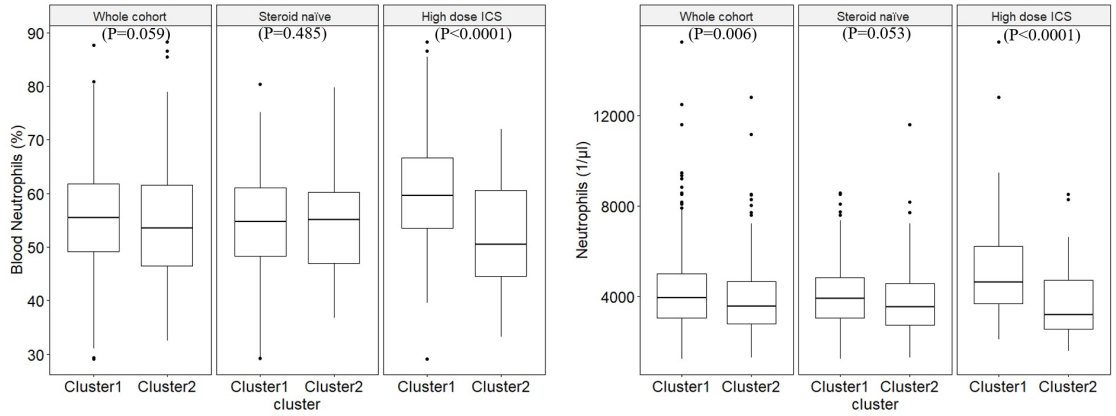

Figure 7.7: Blood Neutrophils for three categories of cohorts in two clusters

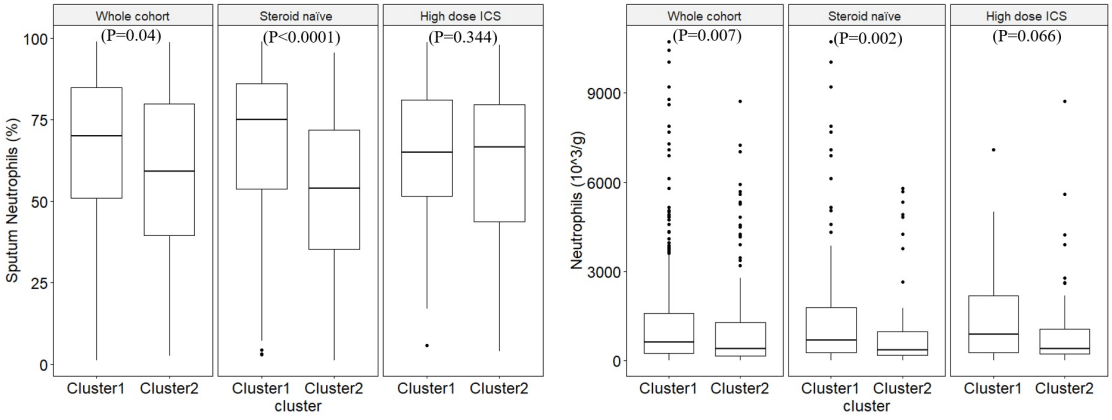

Figure 7.8: Sputum Neutrophils for three categories of cohorts in two clusters

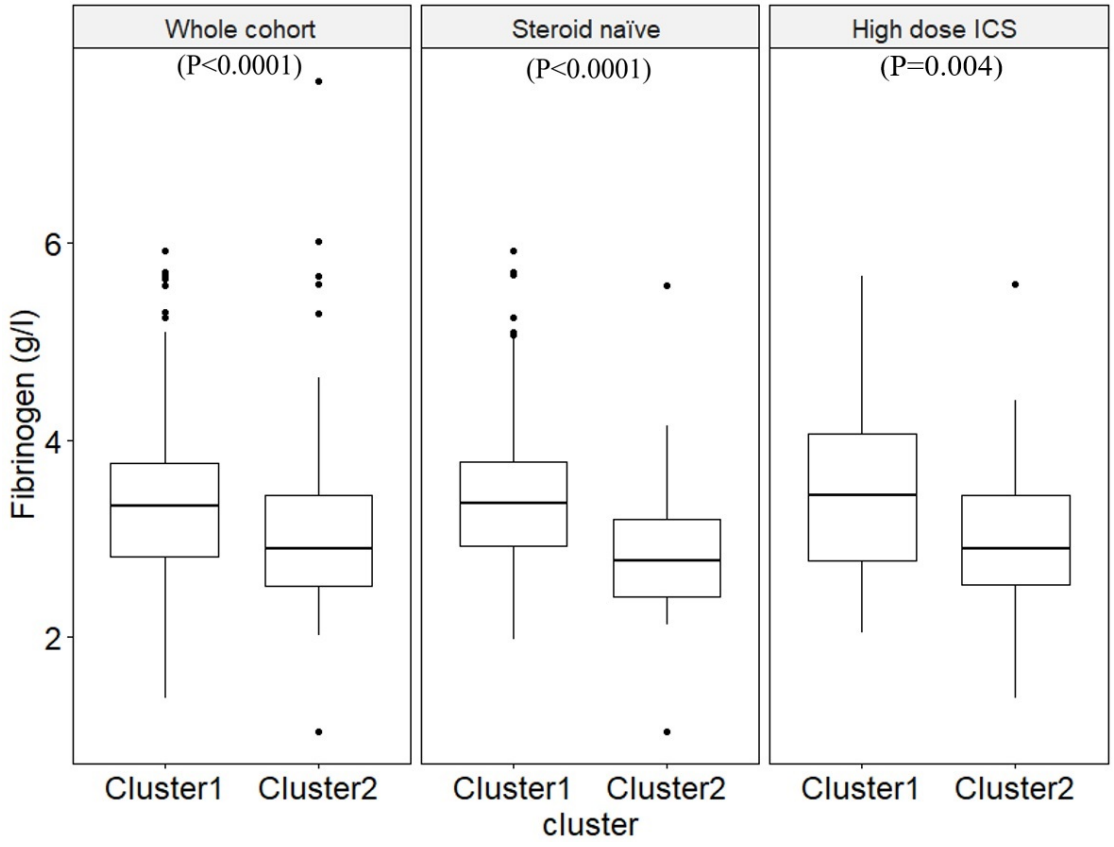

Figure 7.9: Fibrinogen (g/l) for three categories of cohorts in two clusters

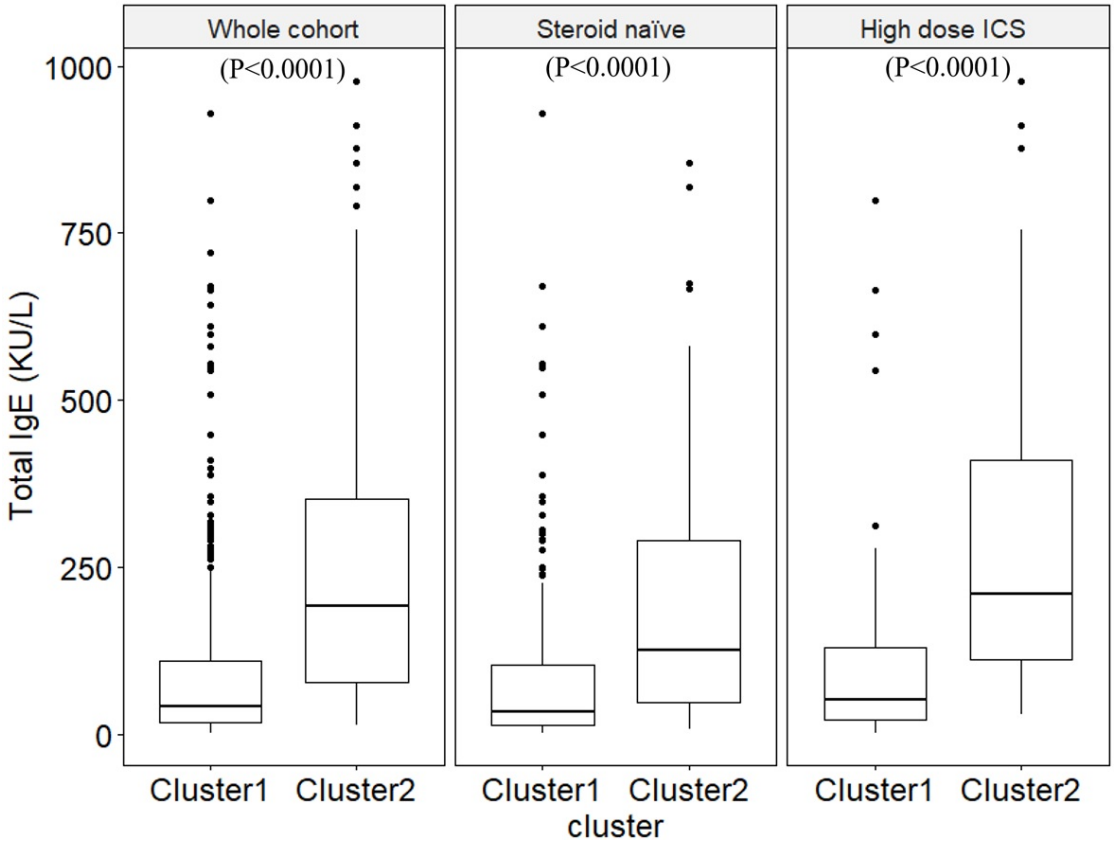

Figure 7.10: Total IgE(KU/L) for three categories of cohorts in two clusters

CHAPTER 8

Discussion and Conclusion

COPD and asthma are the most common airway obstructive diseases affecting 15-20% of the population in western countries and accounting for substantial costs in terms of public health expenditure. It is now well accepted that the terms COPD and asthma actually hide considerable heterogeneity among the patients.

The current definition of COPD is based on a spirometric abnormality (i.e. post bronchodilation FEV1/FVC < 70%) arising from chronic exposure to toxic gas or particulate matters among which tobacco smoke is playing a critical role. Patients satisfying this criterion may actually be very heterogeneous in terms of associated comorbidities, emphysema, airway, and systemic inflammation, symptomatic expression and exacerbation rates, and risk of mortality. Similarly, the current definition of asthma, which implies excessive fluctuation of airway caliber over short period of times allows patients with very different inflammatory, functional and symptomatic profiles to be classified as asthmatics. In particular, it has become clear that eosinophilic and non-eosinophilic asthma phenotypes may behave differently in terms of response to pharmacological treatment. Although there was a time when clinicians used the concept of "one size fits all" to treat chronic airway diseases with ICS combined LABA being prescribed indistinctively to every patient, things are rapidly changing to follow the path of precision medicine leading to more personalized treatments. This may have advantages both in terms of treatment efficiency but also in terms of drug cost sparing (Louis and Roche, 2017).

Cluster analysis has become a popular statistical method to identify phenotypes among a large set of patients. The cluster analysis is a well-known unsupervised learning methodology that creates homogeneous groups based on multiple variables. Once the original cohorts have been selected, missing values and the set of variables of interest are the first initial considerations for cluster analysis. Therefore, cluster analysis on incomplete datasets with multidimensional variables are unavoidable challenges in this thesis. However, there are a number of studies and methods available in the literature that apply cluster analysis to COPD and asthma patients (Horne et al., 2020). There is no clear indication as to which method of cluster analysis on incomplete datasets with multidimensional variables is more efficient or effective than the other. In the literature on clustering, the most common method of handling missing values is the complete case analysis by excluding all patients with any unrecorded values. The majority of these studies considered mixed-type variables. However, the variable selection methods were preferred to the variable reduction methods in these studies (Horne et al., 2020).

The major objective of this thesis was to determine homogeneous groups in chronic obstructive pulmonary disease, eosinophilic asthmatic patients, and non-eosinophilic asthmatic patients. As a result, this thesis provides original data on cohorts of COPD and eosinophilic and non-eosinophilic asthmatics, indicating substantial heterogeneity between clusters and, in asthma, highlighting the differences that may exist inside prespecified airway inflammatory phenotypes. On a statistical point of view and after considering all drawbacks and comprehensively evaluating the cluster analysis methods, we proposed in this thesis a new competitive and complex statistical analysis framework to combine the efficient methods for clustering the datasets containing a large number of variables with missing values. Therefore, a proposed framework for handling missing values using multiple imputation and variable reduction in cluster analysis which contains a new method based on mixture multivariate multinomial model (4M) was introduced. In this perspective, comprehensive simulation scenarios were designed to evaluate the performance of the proposed framework with parallel frameworks and methods using simulated datasets with known clustering results under different missingness and overlapping rates. Due to these scenarios on the simulated datasets, this framework was found to be effective in cluster analysis using multiple imputation when applied to multidimensional incomplete data and has a higher performance than competing methods.

As explained, we combined methods from handling missing data, dimension reduction, cluster analysis, and consensus clustering to discover homogenous clusters in multidimensional incomplete data. In terms of the objects contain by missing values, we have stressed the importance of applying multiple imputation to consider the uncertainty in the analysis. We attempted to present how the missing data influence the different results of a cluster analysis. furthermore, in the literature, many methods are introduced for applying dimension reduction and cluster analysis. Some of these methods work better under some research questions or some specific properties of the datasets. There is no priority in methods of cluster analysis and variable reduction and no clear guidelines as to which ones should be used. However, in this thesis, the well-known and practical methods were introduced and compared. The proposed framework is very flexible and allows users to replace these methods and apply their own methods for clustering and variable reduction. Consequently, it is important to note that the proposed framework can be adapted to replace cluster analysis and dimension reduction methods with alternative methods.

The biggest advantage of this thesis, despite all of the challenges, is applying multiple imputation in cluster analysis, and the consequence of applying multiple imputation is proposing a new framework based on Rubin's rule for cluster analysis in multidimensional incomplete datasets and a new method for consensus clustering.

In multidimensional datasets, the complete case study leads to biased results and lost efficiency, especially in our datasets that should exclude a high percentage of the objects containing at least one missing value. In this thesis, our concern was on the uncertainty in missing values and, consequently, multiple imputation was applied. However, single imputation is another method for handling missing values. In this method, bias is reduced by using an appropriate method of single imputation. However, this method does not account for the uncertainty in the dataset. We believe this method has an effect on the final clustering results so that single imputation and the efficient methods of the single imputation could be a topic for the next comparisons.

We performed our first cluster analysis on a selected population of COPD. We took care of selecting what could be considered "pure COPD" avoiding the population that mixes a previous history of asthma before the age of 40 years and the development of COPD later in life as a consequence of smoking. By applying the framework, three different clusters, which shared the functional definition of COPD and a similar smoking history. The three clusters differed by the sex ratio, the lung function

impairment and the extent of granulocytic airway inflammation, and the propensity to exacerbation but interestingly not by the symptomatic expression as reflected by the CAT score. This further highlights the importance of going beyond symptom collection when assessing a disease like COPD. T2 biomarkers such as FeNO, blood eosinophils, and serum IgE were not different between the clusters whereas absolute sputum eosinophils were increased together with sputum neutrophils in the cluster that showed the most severe clinical outcomes. Remarkably, FeNO levels, atopic status as well as airway colonization by potentially pathogenic microorganisms were low across the three clusters. Overall our study points to a considerable heterogeneity among stable COPD patients recruited from ambulatory care even if they deny any previous history of asthma. The summarization of these clusters is shown in Figure 8.1.

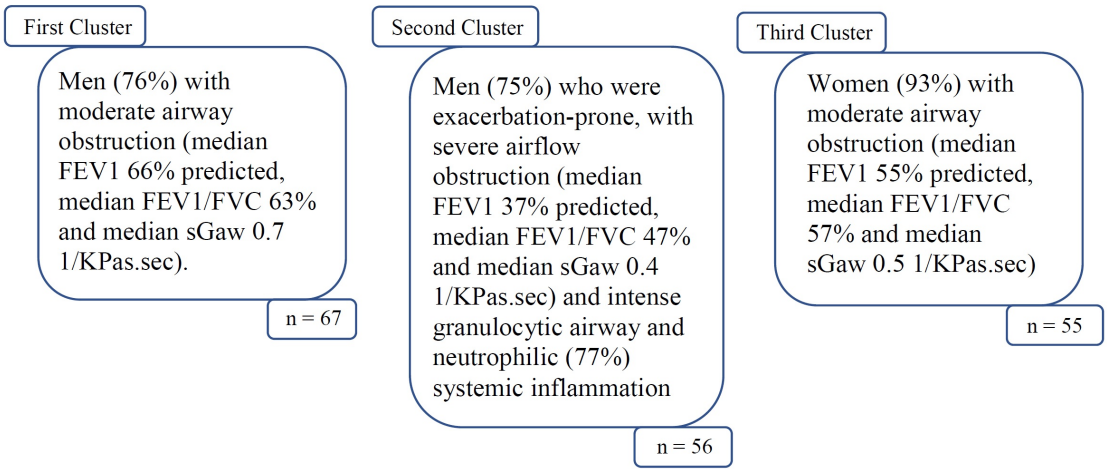

Figure 8.1: General interpretation for COPD clustering (n = 178 patients)

Although the retrospective nature of our studies was a limitation, the strengths were that asthma was carefully diagnosed according to recognized functional criteria and that we collected sputum samples in a large number of patients allowing for airway inflammation assessment. In our eosinophilic and non-eosinophilic asthma studies, the clustering framework yielded two clusters mainly structured by age, atopic status, the intensity of granulocytic airway inflammation, and magnitude of lung function impairment. To the best of our knowledge, there has been no cluster analysis study specifically focused on eosinophilic and non-eosinophilic asthma in the past.

The interest in focusing on eosinophilic asthma was that previous a study sug-

gested that eosinophils found in the respiratory tract may actually differ in their function some of them being clearly pro-inflammatory while others may a regulatory role in dampening airway inflammation (Mesnil et al., 2016). Our data brings clear and novel results with two clusters among eosinophilic asthmatics. One cluster included a majority of non-atopic asthmatics and showed corticoresistance with severe lung function impairment and poor asthma control. The other cluster, which was dominant in terms of number of patients, included a large majority of atopic patients with no functional impairment and showed a relatively good asthma control with a low treatment burden.

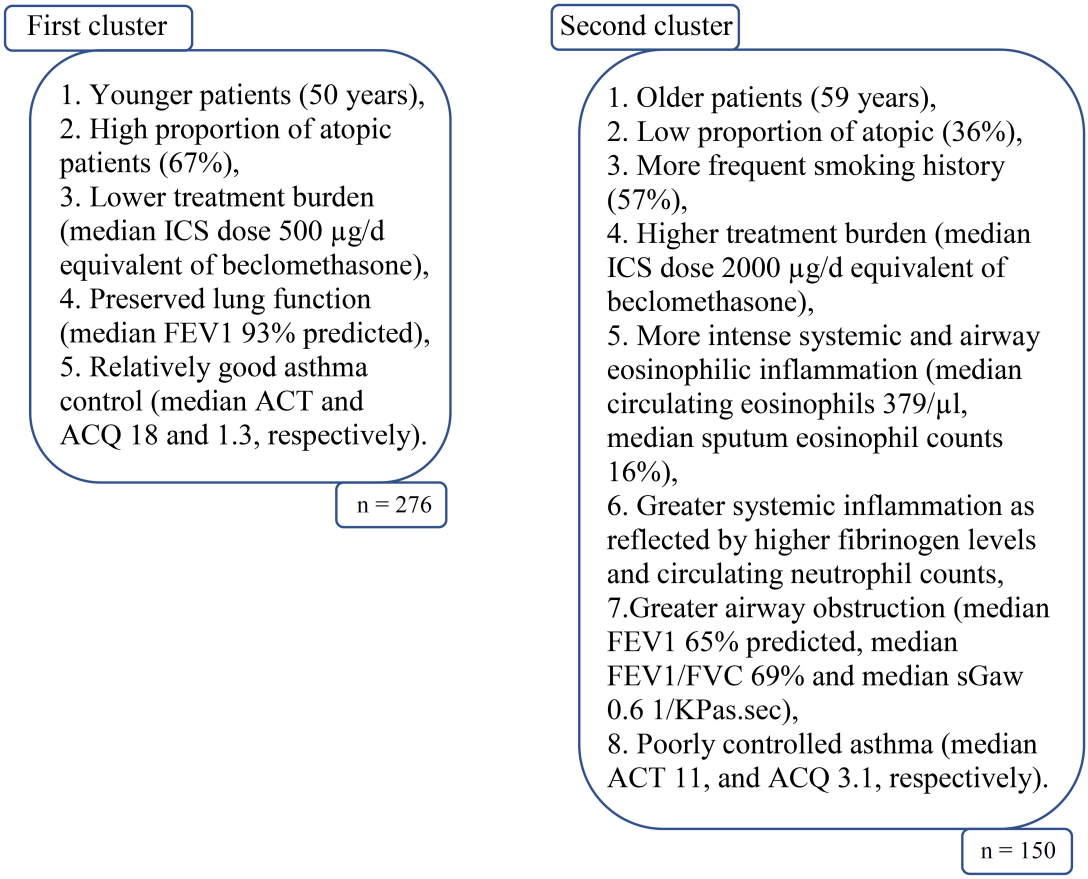

Figure 8.2: General interpretation for eosinophilic asthmatic patients clustering (n = 426 patients)

Even if the magnitude of airway eosinophilic inflammation was more pronounced in the non-atopic cluster our finding would support the idea that eosinophils may indeed be different in the two clusters and that significant eosinophilic airway infiltration may coexist with mild asthma. In addition, by showing higher cortisol levels in the most clinically severe cluster among those who are treated with high dose ICS

we hint towards a true systemic corticoresistance in this cluster, which extends what was already been demonstrated at a local level in the airways. The patients from this cluster are known to greatly benefit from treatment targeting interleukine-5 or interleukine-4/IL-13 (Brusselle and Koppelman, 2022). The results of these analyses are summarized in Figure 8.2.

The study on non-eosinophilic asthma comprised a large asthmatic population highlighting the numerical importance of this population. Our results identified two clusters, one featuring a frequent, but not uniform, smoking history and showing a dominant airway neutrophilic inflammation, the other featuring an almost exclusive atopic cluster with a younger age and no significant airway granulocytic inflammation, entering the category of paucigranulocytic asthma. Exposure to tobacco, air pollutants or microbes may contribute to the neutrophilic airway inflammation and asthma symptoms and play a major role in the first cluster (Fitzpatrick et al., 2020). Atopic asthmatics without eosinophilia may reflect the lack of exposure to the allergen to which the patient is sensitized. We also know that some patients may present intermittent eosinophilia, likely to follow the extend of allergen exposure (McGrath et al., 2012).

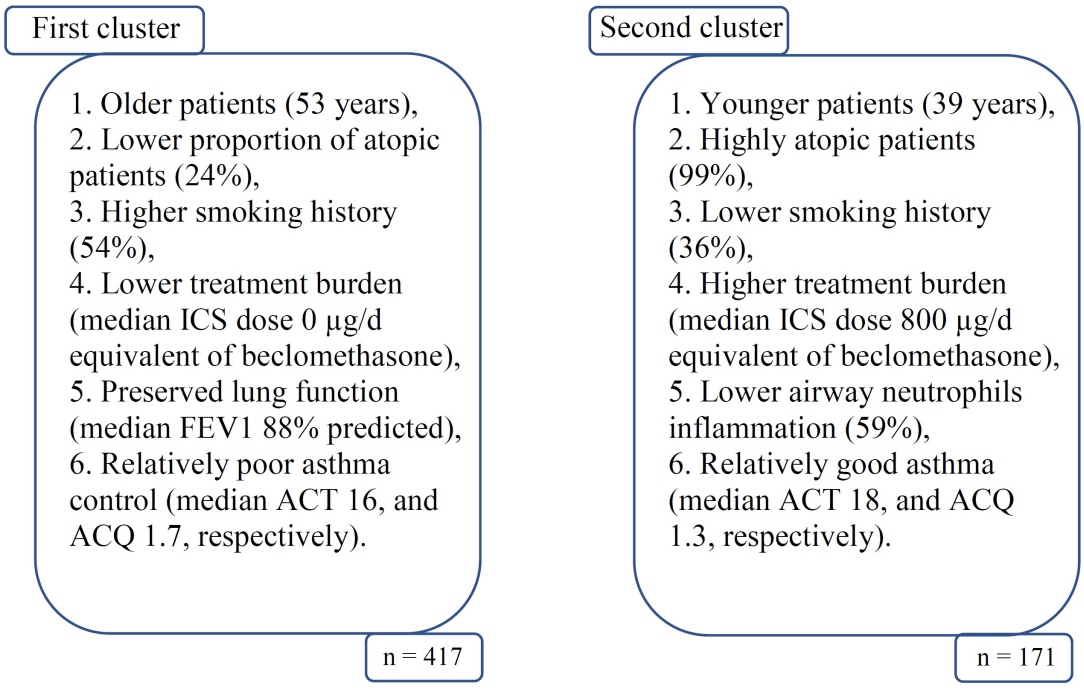

Figure 8.3: General interpretation for non-eosinophilic asthmatic patients clustering (n = 588 patients)

Why patients with paucigranulocytic asthma may still experience symptoms and poor asthma control is an interesting question but there are arguments in the literature to suggest that asthma harbor a fundamental smooth muscle abnormality the contraction of which may be triggered in the absence of overt airway inflammation (An et al., 2007). The relatively high frequency of non-eosinophilic asthma raises a real question of public health as this population was generally found to be poorly responsive to treatment with ICS (Lazarus et al., 2019), and this is particularly the case of the patients displaying marked airway neutrophilia (Green, 2002). Of course, some of the non-eosinophilic asthmatics already treated by ICS are native eosinophilic patient the inflammation of whom has been controlled by the ICS and show a rise in sputum eosinophils when stepping down ICS (Demarche et al., 2018). A summary of the results of these analyses is shown in Figure 8.3.

In this thesis, we have focused on retrospective cross-sectional studies for mixed, continuous, and categorical, variables. As a future direction, it would be interesting to consider the effect of time in our study and clustering the objects during the time. We suggest applying the framework presented in this thesis for handling missing values and consensus clustering, the methods for variable reduction and cluster analysis can be replaced with proper methods for longitudinal data. A longitudinal clustering approach provides a detailed and comprehensive description of objects' time profiles. In this study, we applied the hard clustering method in which each object is only assigned to one cluster. We propose for future studies to consider the soft clustering which does not consider the clustering results as a binary solution and grouping the objects such that an object can exist in multiple clusters with specific clustering probability. These probabilities are ranging from 0 to 1 and indicate how similar an object is to the mean of the cluster.

On the medical side, the clinical validity of our clusters should be validated in other cohorts and most importantly in a longitudinal study as the interest of clustering mainly resides in finding groups of patients that may share common prognosis and treatment response over time. Of particular interest would be to assess the hospitalization rate and the mortality over a five years period in our COPD clusters and the propensity to lung function decline and exacerbation in the same time frame in our asthma clusters.

Bibliography

Abdel-Aziz, M. I., P. Brinkman, S. J. Vijverberg, A. H. Neerincx, J. H. Riley, S. Bates, S. Hashimoto, N. Z. Kermani, K. F. Chung, R. Djukanovic, S.-E. Dahlén, I. M. Adcock, P. H. Howarth, P. J. Sterk, A. D. Kraneveld, and A. H. Maitland-van der Zee (2021, January). Sputum microbiome profiles identify severe asthma phenotypes of relative stability at 12 to 18 months. *Journal of Allergy and Clinical Immunology 147*(1), 123–134.

Agusti, A., E. Bel, M. Thomas, C. Vogelmeier, G. Brusselle, S. Holgate, M. Humbert, P. Jones, P. G. Gibson, J. Vestbo, R. Beasley, and I. D. Pavord (2016, February). Treatable traits: toward precision medicine of chronic airway diseases. *European Respiratory Journal 47*(2), 410–419.

Altenburg, W. A., M. H. de Greef, N. H. ten Hacken, and J. B. Wempe (2012, May). A better response in exercise capacity after pulmonary rehabilitation in more severe COPD patients. *Respiratory Medicine 106*(5), 694–700.

Amaral, A. F. S., R. B. Newson, M. J. Abramson, J. M. Antó, R. Bono, A. G. Corsico, R. de Marco, P. Demoly, B. Forsberg, T. Gislason, J. Heinrich, I. Huerta, C. Janson, R. Jôgi, J.-L. Kim, J. Maldonado, J. Martinez-Moratalla Rovira, C. Neukirch, D. Nowak, I. Pin, N. Probst Hensch, C. Raherison Semjen, C. Svanes, I. Urrutia Landa, R. van Ree, S. A. Versteeg, J. Weyler, J.-P. Zock, P. G. J. Burney, and D. L. Jarvis (2016). Changes in IgE sensitization and total IgE levels over 20 years of follow-up. *The Journal of Allergy and Clinical Immunology 137*(6), 1788–1795.e9.

An, S. S., T. R. Bai, J. H. T. Bates, J. L. Black, R. H. Brown, V. Brusasco, P. Chitano,

L. Deng, M. Dowell, D. H. Eidelman, B. Fabry, N. J. Fairbank, L. E. Ford, J. J. Fredberg, W. T. Gerthoffer, S. H. Gilbert, R. Gosens, S. J. Gunst, A. J. Halayko, R. H. Ingram, C. G. Irvin, A. L. James, L. J. Janssen, G. G. King, D. A. Knight, A. M. Lauzon, O. J. Lakser, M. S. Ludwig, K. R. Lutchen, G. N. Maksym, J. G. Martin, T. Mauad, B. E. McParland, S. M. Mijailovich, H. W. Mitchell, R. W. Mitchell, W. Mitzner, T. M. Murphy, P. D. Pare, R. Pellegrino, M. J. Sanderson, R. R. Schellenberg, C. Y. Seow, P. S. P. Silveira, P. G. Smith, J. Solway, N. L. Stephens, P. J. Sterk, A. G. Stewart, D. D. Tang, R. S. Tepper, T. Tran, and L. Wang (2007). Airway smooth muscle dynamics: a common pathway of airway obstruction in asthma. *European Respiratory Journal 29*(5), 834–860.

Basagaña, X., J. Barrera-Gómez, M. Benet, J. M. Antó, and J. Garcia-Aymerich (2013, April). A framework for multiple imputation in cluster analysis. *American Journal of Epidemiology 177*(7), 718–725.

Bettiol, J., M. Radermecker, J. Sele, M. Henquet, D. Cataldo, and R. Louis (1999, November). Airway mast-cell activation in asthmatics is associated with selective sputum eosinophilia. *Allergy 54*(11), 1188–1193.

Bleecker, E. R., J. M. FitzGerald, P. Chanez, A. Papi, S. F. Weinstein, P. Barker, S. Sproule, G. Gilmartin, M. Aurivillius, V. Werkström, and M. Goldman (2016, October). Efficacy and safety of benralizumab for patients with severe asthma uncontrolled with high-dosage inhaled corticosteroids and long-acting $\beta 2-$agonists (SIROCCO): a randomised, multicentre, placebo-controlled phase 3 trial. *The Lancet 388*(10056), 2115–2127.

Bourdin, A. and P. Chanez (2013, June). Clustering in asthma: why, how and for how long? *European Respiratory Journal 41*(6), 1247–1248.

Brightling, C. E. (2006, March). Sputum Induction in Asthma. *Chest 129*(3), 503–504.

Brightling, C. E., P. Nair, D. J. Cousins, R. Louis, and D. Singh (2021, October). Risankizumab in Severe Asthma – A Phase 2a, Placebo-Controlled Trial. *New England Journal of Medicine 385*(18), 1669–1679.

Bruckers, L. (2014). *Challenges in Cluster Analyses for Longitudinal Data.* Ph. D. thesis, Hasselt University.

Bruckers, L., G. Molenberghs, and P. Dendale (2017, September). Clustering multiply imputed multivariate high-dimensional longitudinal profiles: Clustering multiply imputed data. *Biometrical Journal 59*(5), 998–1015. Number: 5.

Bruckers, L., G. Molenberghs, B. Pulinx, F. Hellenthal, and G. Schurink (2018). Cluster analysis for repeated data with dropout: Sensitivity analysis using a distal event. *Journal of Biopharmaceutical Statistics 28*(5), 983–1004.

Brusselle, G. G. and G. H. Koppelman (2022, January). Biologic therapies for severe asthma. *New England Journal of Medicine 386*(2), 157–171.

Burgel, P.-R., J.-L. Paillasseur, D. Caillaud, I. Tillie-Leblond, P. Chanez, R. Escamilla, I. Court-Fortune, T. Perez, P. Carre, N. Roche, and on behalf of the Initiatives BPCO Scientific Committee (2010, September). Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *European Respiratory Journal 36*(3), 531–539.

Burgel, P.-R., J.-L. Paillasseur, W. Janssens, J. Piquet, G. Ter Riet, J. Garcia-Aymerich, B. Cosio, P. Bakke, M. A. Puhan, A. Langhammer, I. Alfageme, P. Almagro, J. Ancochea, B. R. Celli, C. Casanova, J. P. de Torres, M. Decramer, A. Echazarreta, C. Esteban, R. M. Gomez Punter, M. K. Han, A. Johannessen, B. Kaiser, B. Lamprecht, P. Lange, L. Leivseth, J. M. Marin, F. Martin, P. Martinez-Camblor, M. Miravitlles, T. Oga, A. Sofia Ramírez, D. D. Sin, P. Sobradillo, J. J. Soler-Cataluña, A. M. Turner, F. J. Verdu Rivera, J. B. Soriano, N. Roche, and Initiatives BPCO, EABPCO, Leuven and 3CIA study groups (2017). A simple algorithm for the identification of clinical COPD phenotypes. *The European Respiratory Journal 50*(5).

Burgel, P.-R., J.-L. Paillasseur, B. Peene, D. Dusser, N. Roche, J. Coolen, T. Troosters, M. Decramer, and W. Janssens (2012, December). Two Distinct Chronic Obstructive Pulmonary Disease (COPD) Phenotypes Are Associated with High Risk of Mortality. *PLoS ONE 7*(12), e51048.

Burgel, P.-R., N. Roche, J.-L. Paillasseur, I. Tillie-Leblond, P. Chanez, R. Escamilla, I. Court-Fortune, T. Perez, P. Carré, and D. Caillaud (2012, August). Clinical COPD phenotypes identified by cluster analysis: validation with mortality. *European Respiratory Journal 40*(2), 495–496.

Bury, T. B., J. L. Corhay, and M. F. Radermecker (1992, December). Histamine-induced inhibition of neutrophil chemotaxis and T-lymphocyte proliferation in man. *Allergy 47*(6), 624–629.

Calinski, T. and J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods 3*(1), 1–27.

Castaldi, P. J., M. Benet, H. Petersen, N. Rafaels, J. Finigan, M. Paoletti, H. Marike Boezen, J. M. Vonk, R. Bowler, M. Pistolesi, M. A. Puhan, J. Anto, E. Wauters, D. Lambrechts, W. Janssens, F. Bigazzi, G. Camiciottoli, M. H. Cho, C. P. Hersh, K. Barnes, S. Rennard, M. P. Boorgula, J. Dy, N. N. Hansel, J. D. Crapo, Y. Tesfaigzi, A. Agusti, E. K. Silverman, and J. Garcia-Aymerich (2017, November). Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax 72*(11), 998–1006.

Cataldo, D., J.-L. Corhay, E. Derom, R. Louis, E. Marchand, A. Michils, V. Ninane, R. Peché, C. Pilette, W. Vincken, and W. Janssens (2017). A Belgian survey on the diagnosis of asthma-COPD overlap syndrome. *International Journal of Chronic Obstructive Pulmonary Disease 12*, 601–613.

Celli, B. R., N. Locantore, J. Yates, R. Tal-Singer, B. E. Miller, P. Bakke, P. Calverley, H. Coxson, C. Crim, L. D. Edwards, D. A. Lomas, A. Duvoix, W. MacNee, S. Rennard, E. Silverman, J. Vestbo, E. Wouters, A. Agustì, and ECLIPSE Investigators (2012, May). Inflammatory biomarkers improve clinical prediction of mortality in chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine 185*(10), 1065–1072.

Chung, K. F., S. E. Wenzel, J. L. Brozek, A. Bush, M. Castro, P. J. Sterk, I. M. Adcock, E. D. Bateman, E. H. Bel, E. R. Bleecker, L.-P. Boulet, C. Brightling, P. Chanez, S.-E. Dahlen, R. Djukanovic, U. Frey, M. Gaga, P. Gibson, Q. Hamid, N. N. Jajour, T. Mauad, R. L. Sorkness, and W. G. Teague (2014, February). International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *European Respiratory Journal 43*(2), 343–373.

Couillard, S., R. Shrimanker, S. Lemaire-Paquette, G. M. Hynes, C. Borg, C. Connolly, S. J. Thulborn, A. Moran, S. Poole, S. Morgan, T. Powell, I. Pavord, and T. Hinks (2022). Longitudinal changes in sputum and blood inflammatory mediators during FeNO suppression testing. *Thorax*, thoraxjnl–2021–217994.

Delvaux, M. (2004, February). Nebulised salbutamol administered during sputum induction improves bronchoprotection in patients with asthma. *Thorax 59*(2), 111–115.

Demarche, S., F. Schleich, M. Henket, V. Paulus, R. Louis, and T. Van Hees (2018, May). Step-down of inhaled corticosteroids in non-eosinophilic asthma: A prospective trial in real life. *Clinical & Experimental Allergy 48*(5), 525–535.

Demarche, S., F. Schleich, M. Henket, V. Paulus, T. Van Hees, and R. Louis (2016, April). Detailed analysis of sputum and systemic inflammation in asthma phenotypes: are paucigranulocytic asthmatics really non-inflammatory? *BMC pulmonary medicine 16*, 46.

Demarche, S. F., F. N. Schleich, V. A. Paulus, M. A. Henket, T. J. Van Hees, and R. E. Louis (2017, December). Is it possible to claim or refute sputum eosinophils ≥ 3% in asthmatics with sufficient accuracy using biomarkers? *Respiratory Research 18*(1), 133.

Denton, E., D. B. Price, T. N. Tran, G. W. Canonica, A. Menzies-Gow, J. M. FitzGerald, M. Sadatsafavi, L. Perez de Llano, G. Christoff, A. Quinton, C. K. Rhee, G. Brusselle, C. Ulrik, N. Lugogo, F. Hore-Lacy, I. Chaudhry, L. Bulathsinhala, R. B. Murray, V. A. Carter, and M. Hew (2021, July). Cluster Analysis of Inflammatory Biomarker Expression in the International Severe Asthma Registry. *The Journal of Allergy and Clinical Immunology: In Practice 9*(7), 2680–2688.e7.

Douwes, J. (2002, July). Non-eosinophilic asthma: importance and possible mechanisms. *Thorax 57*(7), 643–648.

Dunn, J. C. (1974, January). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics 4*(1), 95–104.

Esteban-Gorgojo, I., D. Antolìn-Amèrigo, J. Domìnguez-Ortega, and S. Quirce (2018, October). Non-eosinophilic asthma: current perspectives. *Journal of Asthma and Allergy Volume 11*, 267–281.

Estivill-Castro, V. (2002, June). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter 4*(1), 65–75.

Everitt, B. (Ed.) (2011). *Cluster analysis* (5th ed ed.). Wiley series in probability and statistics. Chichester, West Sussex, U.K: Wiley. OCLC: ocn666867900.

Fahy, J. V., J. Liu, H. Wong, and H. A. Boushey (1994, June). Analysis of cellular and biochemical constituents of induced sputum after allergen challenge: A method for studying allergic airway inflammation. *Journal of Allergy and Clinical Immunology 93*(6), 1031–1039.

Fens, N., A. G. van Rossum, P. Zanen, B. van Ginneken, R. J. van Klaveren, A. H. Zwinderman, and P. J. Sterk (2013, June). Subphenotypes of Mild-to-Moderate COPD by Factor and Cluster Analysis of Pulmonary Function, CT Imaging and Breathomics

in a Population-Based Survey. *COPD: Journal of Chronic Obstructive Pulmonary Disease 10*(3), 277–285.

Fitzpatrick, A. M., B. E. Chipps, F. Holguin, and P. G. Woodruff (2020, February). T2-Low Asthma: Overview and Management Strategies. *The Journal of Allergy and Clinical Immunology: In Practice 8*(2), 452–463.

Foss, A. H. and M. Markatou (2018). **kamila** : Clustering Mixed-Type Data in *R* and **Hadoop**. *Journal of Statistical Software 83*(13). Number: 13.

Freitas, P. D., R. F. Xavier, V. M. McDonald, P. G. Gibson, L. Cordova-Rivera, K. C. Furlanetto, J. M. de Oliveira, R. M. Carvalho-Pinto, A. Cukier, R. Stelmach, and C. R. F. Carvalho (2020, July). Identification of asthma phenotypes based on extrapulmonary treatable traits. *European Respiratory Journal*, 2000240.

Garcia-Aymerich, J., F. P. Gomez, M. Benet, E. Farrero, X. Basagaña, A. Gayete, C. Pare, X. Freixa, J. Ferrer, A. Ferrer, J. Roca, J. B. Galdiz, J. Sauleda, E. Monso, J. Gea, J. A. Barbera, A. Agusti, J. M. Anto, and on behalf of the PAC-COPD Study Group (2011, May). Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax 66*(5), 430–437.

Gerday, S., F. Schleich, M. Henket, F. Guissard, V. Paulus, and R. Louis (2022). Revisiting differences between atopic and non-atopic asthmatics: When age is shaping airway inflammatory profile. *World Allergy Organization Journal 15*(6), 100655.

Gold, L. S., N. Smith, F. C. Allen-Ramey, R. A. Nathan, and S. D. Sullivan (2012, October). Associations of patient outcomes with level of asthma control. *Annals of Allergy, Asthma & Immunology 109*(4), 260–265.e2.

Gold, L. S., N. Smith, F. C. Allen-Ramey, R. A. Nathan, and S. D. Sullivan (2019, October). Associations of patient outcomes with level of asthma control. *Annals of Allergy, Asthma & Immunology 109*(4), 260–265.e2.

Gordon, A. (1999). *Classification* (2nd Edition ed.). Chapman & Hall/CRC, Boca Raton, Florida.

Gordon, A. D. and M. Vichi (2001, June). Fuzzy partition models for fitting a set of partitions. *Psychometrika 66*(2), 229–247. Number: 2.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics 27*(4), 857–871. Publisher: [Wiley, International Biometric Society].

Graff, S., S. Vanwynsberghe, G. Brusselle, S. Hanon, C. Sohy, L. J. Dupont, R. Peche, A. Michils, C. Pilette, G. Joos, R. E. Louis, and F. N. Schleich (2020, December). Chronic oral corticosteroids use and persistent eosinophilia in severe asthmatics from the Belgian severe asthma registry. *Respiratory Research 21*(1), 214.

Green, R. H. (2002, October). Analysis of induced sputum in adults with asthma: identification of subgroup with isolated sputum neutrophilia and poor response to inhaled corticosteroids. *Thorax 57*(10), 875–879.

Haldar, P., I. D. Pavord, D. E. Shaw, M. A. Berry, M. Thomas, C. E. Brightling, A. J. Wardlaw, and R. H. Green (2008, August). Cluster Analysis and Clinical Asthma Phenotypes. *American Journal of Respiratory and Critical Care Medicine 178*(3), 218–224.

Hamerly, G. J. (2003). *Learning Structure and Concepts in Data using Data Clustering*. Ph. D. thesis, University of California, San Diego.

Han, M. K., A. Agusti, P. M. Calverley, B. R. Celli, G. Criner, J. L. Curtis, L. M. Fabbri, J. G. Goldin, P. W. Jones, W. MacNee, B. J. Make, K. F. Rabe, S. I. Rennard, F. C. Sciurba, E. K. Silverman, J. Vestbo, G. R. Washko, E. F. M. Wouters, and F. J. Martinez (2010, September). Chronic Obstructive Pulmonary Disease Phenotypes: The Future of COPD. *American Journal of Respiratory and Critical Care Medicine 182*(5), 598–604.

Hancox, R. J., D. C. Cowan, R. E. Aldridge, J. O. Cowan, R. Palmay, A. Williamson, G. I. Town, and D. R. Taylor (2012, April). Asthma phenotypes: Consistency of classification using induced sputum: Stability of asthma phenotypes. *Respirology 17*(3), 461–466.

Hartl, S., M.-K. Breyer, O. C. Burghuber, A. Ofenheimer, A. Schrott, M. H. Urban, A. Agusti, M. Studnicka, E. F. Wouters, and R. Breyer-Kohansal (2020, May). Blood eosinophil count in the general population: typical values and potential confounders. *European Respiratory Journal 55*(5), 1901874.

Hastie, A. T., D. T. Mauger, L. C. Denlinger, A. Coverstone, M. Castro, S. Erzurum, N. Jarjour, B. D. Levy, D. A. Meyers, W. C. Moore, B. R. Phillips, S. E. Wenzel, J. V. Fahy, E. Israel, and E. R. Bleecker (2021, April). Mixed Sputum Granulocyte Longitudinal Impact on Lung Function in the Severe Asthma Research Program. *American Journal of Respiratory and Critical Care Medicine 203*(7), 882–892.

Hastie, A. T., W. C. Moore, D. A. Meyers, P. L. Vestal, H. Li, S. P. Peters, and E. R. Bleecker (2010, May). Analyses of asthma severity phenotypes and inflammatory proteins

in subjects stratified by sputum granulocytes. *Journal of Allergy and Clinical Immunology 125*(5), 1028–1036.e13.

Horne, E., H. Tibble, A. Sheikh, and A. Tsanas (2020, May). Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping. *JMIR Medical Informatics 8*(5), e16452.

Hornik, K. (2005). A CLUE for CLUster Ensembles. *Journal of Statistical Software 14*(12). Number: 12.

Hubert, L. and P. Arabie (1985, December). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for clustering data*. Prentice Hall advanced reference series. Englewood Cliffs, N.J: Prentice Hall.

Jatakanon, A., S. Lim, K. Chung, and P. Barnes (1998, November). An inhaled steroid improves markers of airway inflammation in patients with mild asthma. *European Respiratory Journal 12*(5), 1084–1088.

Juniper, E. F., P. M. O'Byrne, G. H. Guyatt, P. J. Ferrie, and D. R. King (1999, October). Development and validation of a questionnaire to measure asthma control. *European Respiratory Journal 14*(4), 902–907. Publisher: European Respiratory Society Section: Original Articles.

Kass, R. E. and A. E. Raftery (1995, June). Bayes Factors. *Journal of the American Statistical Association 90*(430), 773–795. Number: 430 Publisher: Taylor & Francis.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data: an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. New York: Wiley.

Kaur, R. and G. Chupp (2019, July). Phenotypes and endotypes of adult asthma: Moving toward precision medicine. *Journal of Allergy and Clinical Immunology 144*(1), 1–12.

Kuhlen, J. L., A. E. Wahlquist, P. J. Nietert, and S. N. Bains (2014, December). Identification of Asthma Phenotypes in a Tertiary Care Medical Center. *The American Journal of the Medical Sciences 348*(6), 480–485.

Lapperre, T. S., J. B. Snoeck-Stroband, M. M. Gosman, J. Stolk, J. K. Sont, D. F. Jansen, H. A. Kerstjens, D. S. Postma, and P. J. Sterk (2004, September). Dissociation of

Lung Function and Airway Inflammation in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine 170*(5), 499–504.

Lazarus, S. C., J. A. Krishnan, T. S. King, J. E. Lang, K. V. Blake, R. Covar, N. Lugogo, S. Wenzel, V. M. Chinchilli, D. T. Mauger, A.-M. Dyer, H. A. Boushey, J. V. Fahy, P. G. Woodruff, L. B. Bacharier, M. D. Cabana, J. C. Cardet, M. Castro, J. Chmiel, L. Denlinger, E. DiMango, A. M. Fitzpatrick, D. Gentile, A. Hastie, F. Holguin, E. Israel, D. Jackson, M. Kraft, C. LaForce, R. F. Lemanske, F. D. Martinez, W. Moore, W. J. Morgan, J. N. Moy, R. Myers, S. P. Peters, W. Phipatanakul, J. A. Pongracic, L. Que, K. Ross, L. Smith, S. J. Szefler, M. E. Wechsler, and C. A. Sorkness (2019, May). Mometasone or Tiotropium in Mild Asthma with a Low Sputum Eosinophil Level. *New England Journal of Medicine 380*(21), 2009–2019.

Li, T., M. Ogihara, and S. Ma (2010, October). On combining multiple clusterings: an overview and a new perspective. *Applied Intelligence 33*(2), 207–219. Number: 2.

Lim, S., A. Jatakanon, M. John, T. Gilbey, B. O'Connor, K. Chung, and P. Barnes (1999, January). Effect of Inhaled Budesonide on Lung Function and Airway Inflammation: Assessment by Various Inflammatory Markers in Mild Asthma. *American Journal of Respiratory and Critical Care Medicine 159*(1), 22–30.

Louis, R. and N. Roche (2017, September). Personalised medicine: are we ready? *European Respiratory Review 26*(145), 170088.

Louis, R. E. and F. N. Schleich (2021, April). Granulocytic Airway Inflammation and Clinical Asthma Outcomes. *American Journal of Respiratory and Critical Care Medicine 203*(7), 797–799.

MacQueen, J. (1967, January). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics 5.1*, 281–298. Publisher: University of California Press.

Mahalanbois, P. C. On the generalized distance in statistics. *In: Naitional Institute of Sciences of India*, 49–55.

Manise, M., B. Bakayoko, F. Schleich, J.-L. Corhay, and R. Louis (2016, July). IgE mediated sensitisation to aeroallergens in an asthmatic cohort: relationship with inflammatory phenotypes and disease severity. *International Journal of Clinical Practice 70*(7), 596–605.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to information retrieval*. New York: Cambridge University Press. OCLC: ocn190786122.

McGrath, K. W., N. Icitovic, H. A. Boushey, S. C. Lazarus, E. R. Sutherland, V. M. Chinchilli, and J. V. Fahy (2012, March). A Large Subgroup of Mild-to-Moderate Asthma Is Persistently Noneosinophilic. *American Journal of Respiratory and Critical Care Medicine 185*(6), 612–619.

McLachlan, G. J. and D. Peel (2000). *Finite mixture models*. Wiley series in probability and statistics. Applied probability and statistics section. New York: Wiley.

Mesnil, C., S. Raulier, G. Paulissen, X. Xiao, M. A. Birrell, D. Pirottin, T. Janss, P. Starkl, E. Ramery, M. Henket, F. N. Schleich, M. Radermecker, K. Thielemans, L. Gillet, M. Thiry, M. G. Belvisi, R. Louis, C. Desmet, T. Marichal, and F. Bureau (2016, August). Lung-resident eosinophils represent a distinct regulatory eosinophil subset. *Journal of Clinical Investigation 126*(9), 3279–3295.

Miravitlles, M., J. J. Soler-Cataluña, M. Calle, and J. B. Soriano (2013, June). Treatment of COPD by clinical phenotypes: putting old evidence into clinical practice. *European Respiratory Journal 41*(6), 1252–1256.

Miravitlles, M., H. Worth, J. J. Soler-Cataluña, D. Price, F. De Benedetto, N. Roche, N. S. Godtfredsen, T. van der Molen, C.-G. Löfdahl, L. Padullés, and A. Ribera (2016, September). The Relationship Between 24-Hour Symptoms and COPD Exacerbations and Healthcare Resource Use: Results from an Observational Study (ASSESS). *COPD: Journal of Chronic Obstructive Pulmonary Disease 13*(5), 561–568.

Moermans, C., V. Heinen, M. Nguyen, M. Henket, J. Sele, M. Manise, J. L. Corhay, and R. Louis (2011, November). Local and systemic cellular inflammation and cytokine release in chronic obstructive pulmonary disease. *Cytokine 56*(2), 298–304.

Molenberghs, G. and M. Kenward (2007, April). *Missing Data in Clinical Studies*. John Wiley & Sons. Google-Books-ID: SuPJkcfdn1YC.

Moore, W. C., D. A. Meyers, S. E. Wenzel, W. G. Teague, H. Li, X. Li, R. D'Agostino, M. Castro, D. Curran-Everett, A. M. Fitzpatrick, B. Gaston, N. N. Jarjour, R. Sorkness, W. J. Calhoun, K. F. Chung, S. A. A. Comhair, R. A. Dweik, E. Israel, S. P. Peters, W. W. Busse, S. C. Erzurum, and E. R. Bleecker (2010, February). Identification of Asthma Phenotypes Using Cluster Analysis in the Severe Asthma Research Program. *American Journal of Respiratory and Critical Care Medicine 181*(4), 315–323.

Nathan, R. A., C. A. Sorkness, M. Kosinski, M. Schatz, J. T. Li, P. Marcus, J. J. Murray, and T. B. Pendergraft (2004, January). Development of the asthma control testâA survey for assessing asthma control. *Journal of Allergy and Clinical Immunology 113*(1), 59–65.

Newby, C., L. G. Heaney, A. Menzies-Gow, R. M. Niven, A. Mansur, C. Bucknall, R. Chaudhuri, J. Thompson, P. Burton, C. Brightling, and on behalf of the British Thoracic Society Severe Refractory Asthma Network (2014, July). Statistical Cluster Analysis of the British Thoracic Society Severe Refractory Asthma Registry: Clinical Outcomes and Phenotype Stability. *PLoS ONE 9*(7), e102987.

Nunes, C., A. M. Pereira, and M. Morais-Almeida (2017, December). Asthma costs and social impact. *Asthma Research and Practice 3*(1), 1.

O'Byrne, P., L. M. Fabbri, I. D. Pavord, A. Papi, S. Petruzzelli, and P. Lange (2019, July). Asthma progression and mortality: the role of inhaled corticosteroids. *European Respiratory Journal 54*(1), 1900491.

O'Connor, G. T., D. Sparrow, and S. T. Weiss (1989, July). The role of allergy and nonspecific airway hyperresponsiveness in the pathogenesis of chronic obstructive pulmonary disease. *The American Review of Respiratory Disease 140*(1), 225–252.

Paoletti, M., G. Camiciottoli, E. Meoni, F. Bigazzi, L. Cestelli, M. Pistolesi, and C. Marchesi (2009, December). Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes. *Journal of Biomedical Informatics 42*(6), 1013–1021.

Pavord, I. D., R. Beasley, A. Agusti, G. P. Anderson, E. Bel, G. Brusselle, P. Cullinan, A. Custovic, F. M. Ducharme, J. V. Fahy, U. Frey, P. Gibson, L. G. Heaney, P. G. Holt, M. Humbert, C. M. Lloyd, G. Marks, F. D. Martinez, P. D. Sly, E. von Mutius, S. Wenzel, H. J. Zar, and A. Bush (2018, January). After asthma: redefining airways diseases. *The Lancet 391*(10118), 350–400.

Pavord, I. D., S. Korn, P. Howarth, E. R. Bleecker, R. Buhl, O. N. Keene, H. Ortega, and P. Chanez (2012, August). Mepolizumab for severe eosinophilic asthma (DREAM): a multicentre, double-blind, placebo-controlled trial. *380*(9842), 651–659.

Pavord, I. D., S. Lettis, N. Locantore, S. Pascoe, P. W. Jones, J. A. Wedzicha, and N. C. Barnes (2016, February). Blood eosinophils and inhaled corticosteroid/long-acting β2 agonist efficacy in COPD. *Thorax 71*(2), 118–125.

Peters, M. C., S. Kerr, E. M. Dunican, P. G. Woodruff, M. L. Fajt, B. D. Levy, E. Israel, B. R. Phillips, D. T. Mauger, S. A. Comhair, S. C. Erzurum, M. W. Johansson, N. N. Jarjour, A. M. Coverstone, M. Castro, A. T. Hastie, E. R. Bleecker, S. E. Wenzel, and J. V. Fahy (2019, January). Refractory airway type 2 inflammation in a large subgroup of asthmatic patients treated with inhaled corticosteroids. *143*(1), 104–113.e14.

Piacentini, G. L., L. Martinati, S. Mingoni, and A. L. Boner (1996, May). Influence of allergen avoidance on the eosinophil phase of airway inflammation in children with allergic asthma. *Journal of Allergy and Clinical Immunology 97*(5), 1079–1084.

Pistolesi, M., G. Camiciottoli, M. Paoletti, C. Marmai, F. Lavorini, E. Meoni, C. Marchesi, and C. Giuntini (2008, March). Identification of a predominant COPD phenotype in clinical practice. *Respiratory Medicine 102*(3), 367–376.

Quaedvlieg, V., J. Sele, M. Henket, and R. Louis (2009, December). Association between asthma control and bronchial hyperresponsiveness and airways inflammation: a cross-sectional study in daily practice. *Clinical & Experimental Allergy 39*(12), 1822–1829.

Rabe, K. F. (2004, December). Outcome measures in COPD. *Primary Care Respiratory Journal 13*(4), 177–178.

Radermecker, C., R. Louis, F. Bureau, and T. Marichal (2018, October). Role of neutrophils in allergic asthma. *Current Opinion in Immunology 54*, 28–34.

Rand, W. M. (1971, December). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association 66*(336), 846–850. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1971.10482356.

Reddel, H. K., J. Vestbo, A. Agustì, G. P. Anderson, A. T. Bansal, R. Beasley, E. H. Bel, C. Janson, B. Make, I. D. Pavord, D. Price, E. Rapsomaniki, N. Karlsson, D. K. Finch, J. Nuevo, A. de Giorgio-Miller, M. Alacqua, R. Hughes, H. Müllerovà, M. Gerhardsson de Verdier, and for the NOVELTY study investigators (2021, September). Heterogeneity within and between physician-diagnosed asthma and/or COPD: NOVELTY cohort. *European Respiratory Journal 58*(3), 2003927.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. Wiley classics library. Hoboken, N.J: Wiley-Interscience.

Santos, M. S., R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu (2019). Generating Synthetic Missing Data: A Review by Missing Mechanism. *IEEE Access 7*, 11651–11667.

Schleich, F. N., A. Chevremont, V. Paulus, M. Henket, M. Manise, L. Seidel, and R. Louis (2014, July). Importance of concomitant local and systemic eosinophilia in uncontrolled asthma. *The European Respiratory Journal 44*(1), 97–108.

Schleich, F. N., M. Manise, J. Sele, M. Henket, L. Seidel, and R. Louis (2013, December). Distribution of sputum cellular phenotype in a large asthma cohort: predicting factors for eosinophilic vs neutrophilic inflammation. *BMC Pulmonary Medicine 13*(1), 11.

Schwarz, G. (1978, March). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461–464. Number: 2.

Simpson, J. L., R. Scott, M. J. Boyle, and P. G. Gibson (2006, January). Inflammatory subtypes in asthma: Assessment and identification using induced sputum. *Respirology 11*(1), 54–61.

Snider, G. L. (1989, September). Chronic Obstructive Pulmonary Disease: A Definition and Implications of Structural Determinants of Airflow Obstruction for Epidemiology. *American Review of Respiratory Disease 140*(3_pt_2), S3–S8.

Tay, T. R. and M. Hew (2018, July). Comorbid treatable traits in difficult asthma: Current evidence and clinical evaluation. *Allergy 73*(7), 1369–1382.

Thomas, N. (1999, March). Relationships between age, dehydro-epiandrosterone sulphate and plasma glucose in healthy men. *Age and Ageing 28*(2), 217–220.

Thomson, N. (2004, November). Asthma and cigarette smoking. *European Respiratory Journal 24*(5), 822–833.

Thomson, N. C. (2017, May). Asthma and smoking-induced airway disease without spirometric COPD. *European Respiratory Journal 49*(5), 1602061.

Topchy, A., A. K. Jain, and W. Punch (2004, April). A Mixture Model for Clustering Ensembles. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pp. 379–390. Society for Industrial and Applied Mathematics.

van Bragt, J. J., I. M. Adcock, E. H. Bel, G.-J. Braunstahl, A. ten Brinke, J. Busby, G. W. Canonica, H. Cao, K. F. Chung, Z. Csoma, B. Dahlén, E. Davin, S. Hansen, E. Heffler, I. Horvath, S. Korn, M. Kots, P. Kuna, N. Kwon, R. Louis, V. Plaza, C. Porsbjerg, D. Ramos-Barbon, L. B. Richards, S. Škrgat, J. K. Sont, S. J. Vijverberg, E. J. Weersink, V. Yasinska, S. S. Wagers, R. Djukanovic, and A. H. Maitland-van der Zee (2020, January). Characteristics and treatment regimens across ERS SHARP severe asthma registries. *European Respiratory Journal 55*(1), 1901163.

van Buuren, S. (2018, July). *Flexible Imputation of Missing Data, Second Edition* (2 ed.). Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC.

Vos, T. and et al. (2020, October). Global burden of 369 diseases and injuries in 204 countries and territories, 1990 – 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet 396*(10258), 1204–1222.

Wark, P., N. Saltos, J. Simpson, S. Slater, M. Hensley, and P. Gibson (2000, December). Induced sputum eosinophils and neutrophils and bronchiectasis severity in allergic bronchopulmonary aspergillosis. *European Respiratory Journal 16*(6), 1095–1101.

Wenzel, S. E. (2012, May). Asthma phenotypes: the evolution from clinical to molecular approaches. *Nature Medicine 18*(5), 716–725.

Wenzel, S. E. (2016, July). Emergence of Biomolecular Pathways to Define Novel Asthma Phenotypes. Type-2 Immunity and Beyond. *American Journal of Respiratory Cell and Molecular Biology 55*(1), 1–4.

Xu, I., M. Boulay, M. Bertrand, A. Côté, and L. Boulet (2021, June). Comparative features of eosinophilic and non-eosinophilic asthma. *Clinical & Experimental Allergy*, cea.13959.

Yeung, K. Y., D. R. Haynor, and W. L. Ruzzo (2001, April). Validating clustering for gene expression data. *Bioinformatics 17*(4), 309–318.

Zhong, S. and J. Ghosh (2003). A unified framework for model-based clustering. *The Journal of Machine Learning Research 4*, 1001–1037.