A Statistical Framework for Modeling Asthma and COPD Biological Heterogeneity, and a Novel Variable Selection Method for Model-based Clustering

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Michael A Ghebre
M.Sc. (London School of Hygiene and Tropical Medicine) 2010

Department of Infection, Immunity and Inflammation
University of Leicester

2016

A Statistical Framework for Modeling Asthma and COPD Biological Heterogeneity, and a Novel Variable Selection Method for Model-based Clustering

Michael A Ghebre, B.Sc., M.Sc.

abstract

This thesis has two main parts. The first part is an application that focuses on the identification of a statistical framework to model the biological heterogeneity of asthma and COPD using sputum cytokines. Clustering subjects using the actual cytokines measurements may not be straightforward as these mediators have strong correlations, which are currently ignored by standard clustering techniques. Artificial data, which have similar patterns as the cytokines, but with known class membership, are simulated. Several approaches, such as data reduction using factor analysis, were performed on the simulated data to identify suitable representative of the variables and to use as input into clustering algorithm. In the simulation study, using "factor-scores" (derived from factor analysis) as input variables into clustering outperformed the alternative approaches. Thus, this approach was applied to model the biological heterogeneity of asthma and COPD, and identified three stable and three exacerbation clusters, with different proportions of overlap between the diseases.

The second part is a statistical methodology in which a new method for variable selection in model-based clustering was proposed. This method generalizes the approach of Raftery and Dean (2006, JASA 101, 168-178). It relaxes the global prior assumptions of linear-relationships between clustering relevant and irrelevant variables by searching for latent structures among the variables, and accounts for non-linear relationships between these variables by splitting the data into sub-samples. A Gaussian mixture model (unconstrained variance-covariance matrices fitted using the EM-algorithm) is applied to identify the optimal clusters. The new method performed considerably better than the Raftery and Dean technique when applied to simulated and real datasets, and demonstrates that variable selection within clustering can substantially improve the identification of optimal clusters. However, at the moment it perhaps does not perform adequately in uncovering the optimal clusters in the dataset which have strong correlations such as sputum mediators.

Acknowledgment

I would like to express my deepest gratitude to my first supervisor, Prof Chris Brightling, for his excellent supervision, encouragement, and guidance from the initial to the final level, which enabled me to develop an understanding of the subject. He immensely influenced my thinking and research. He is very kind, helpful, and always available whenever needed.

I am very grateful to my second supervisor, Professor John Thompson, who fully supervised the method part of my thesis, and for always being available for further assistance and guidance. Without his great supervision this work would not have been possible. I learned enormously from many enlightening discussions with him and felt privileged to have had opportunities to be supervised by him.

I am also very grateful to my third supervisor, Dr Chris Newby, for his support and encouragement. I am very lucky to have had him as an advisor and colleague over the years. In addition, my thanks go to Prof Paul Burton for his supervision of my first-year PhD till he moved to the University of Bristol, and Dr Richard May (from Medimmune/AstraZeneca) for the enlightening discussions and productive collaboration over the last four years. I would also like to thank Dr Mona Bafadhel, Dr Dhan Desai, Dr Kairabi Haldar and Dr Latifa Chachi, and all those who supported me in any respect during the completion of my PhD.

Sincere thanks and love go to my family including my mum, siblings and friends for their love and prayerful support. I would like to dedicate this thesis to my family Rosi, Noah and Monary, and my mum Rishan Debesay and in memory of my dad Abrha Ghebre who valued education above all.

Contents

List of Figures

List of Tables

Abbreviations

COPD	Chronic obstructive pulmonary disease
GINA	Global Initiative for Asthma
GOLD	Global Initiative for Chronic Obstructive Lung Disease
FEV_1	Forced expiratory volume in the 1^{st} second
FVC	Forced vital capacity
CFU	Colony-forming unit
VAS	Visual analogue scale
ICS	Inhaled Corticosteroids
MSD	Meso scale discovery platform
ELISA	Enzyme linked immunosorbent assay
IL	Interleukin
VEGF	Vascular Endothelial Growth Factor
TNF	Tumour Necrosis Factor
T_H1	T Helper 1
T_H2	T Helper 2
BMI	Body mass index
FS	Factor score
SEM	Standard error of the mean
ANOVA	Analysis of variance
PCA	Principal component analysis
CART	Classification and Regression Trees
LDA	Linear discriminant analysis
LDF	Linear discriminant function
BIC	Bayesian information criterion
E-M algorithm	Expectation maximization algorithm
R&D method	Raftery and Dean method

Thesis Structure and Contributions

Structure of the Thesis

This thesis has two main parts. The first part is an application that focuses on the identification of an appropriate statistical framework for modeling (clustering) the biological heterogeneity of asthma and COPD jointly using sputum cytokines (at both stable and exacerbation states). This part is divided into six interlinked chapters. Chapter 1 covers the general introduction of asthma and COPD (such as definitions, diagnoses, and differences and similarities between the diseases) and future cytokine-based treatments, and application of cluster analysis to model the heterogeneity of the diseases in order to identify novel clusters/subgroups. In chapter 2, the study population and method were described, and explanatory data analysis at stable state was performed on all available demographic, clinical and biological (sputum cytokines) characteristics, and these patterns were compared across asthma and COPD at disease level. In addition, the internal patterns/structures of the cytokines were investigated further (in order to get initial suggestions that which approach could be suitable for modeling the biological heterogeneity of the diseases). In chapter 3, a simulation study was performed in which multivariate data having similar internal patterns/structures as the cytokines, but with known class membership, was simulated. Thenceforth several representatives of the simulated variables were identified, and were independently used as input variables into a clustering algorithm. The methodology for each approach is described, and the corresponding performance and technical limitation/bias is also discussed. In chapter 4, the method that performed best in the simulation study (chapter 3) was applied to asthma and COPD cytokines study to identify stable biological clusters of the diseases. Thereafter the available clinical and biological characteristics are presented across the identified clusters. In chapter 5, the stable biological clusters of asthma and COPD (which were identified in chapter 4) were validated on inde-

pendent asthma and COPD studies using two approaches. However, prior to that further simulation study was performed to investigate the robustness of the validation statistical techniques. Chapter 6 focused on the identification of independent exacerbation biological clusters of asthma and COPD using robust statistical techniques. In addition, the patterns of the microbiome communities at phylum and genus levels were assessed across asthma and COPD, and across the identified exacerbation biological clusters.

The second part is a methodology in which a new variable selection method for model-based clustering is proposed, which generalizes the approach of Raftery and Dean (2006, JASA 101,168-178). This part is presented in chapter 7, and categorised into several sections instead of chapters. In brief, it started with the study objectives and introduction. Then cluster analysis was introduced briefly, and the methodology for model-based clustering (Gaussian mixture model) which is implemented using EM-algorithm is also discussed in detail. In addition, variable selection in cluster analysis is briefed, and variable selection in model-based clustering (how to assess variables for clustering information) and a general overview of Raftery and Dean's method are described in detail. Furthermore, the detailed algorithm of the proposed method is presented; and an instruction how to use the new software in R for the new method is outlined and its performance was assessed using several simulated and real datasets. This part concluded with discussion section in which the advantages and limitations of the new method are discussed.

This thesis concludes in chapter 8, in which its overall contributions, limitations and future direction are discussed. For instance, the biological heterogeneities of asthma and COPD at stable and exacerbation states are described. In addition, the statistical methods which were applied to model the biological heterogeneity of asthma and COPD are briefed, and the future direction to develop into an algorithm is outlined. Furthermore, the performance and future direction of the new method for variable selection in model-based clustering is discussed.

Thesis Contributions

This thesis has several contributions:

- It outlined how to assess the common and distinctive characteristics of asthma and COPD, beyond the way they are categorized in clinic by physicians based on the existing guidelines of the diseases.
- It provided a general statistical framework on how to model (cluster) asthma and COPD subjects based on their correlated cytokines to identify the distinctive and common biological subgroups/clusters.
- It identified a robust statistical approach to validated clusters using new datasets, in a situation where all the variables which are used in the identification of the original clusters may not measured in the new validation dataset.
- Three distinctive asthma and COPD biological subgroups were identified using appropriate statistical techniques, independently at stable and exacerbation states with different proportion of overlap between the two diseases. These subgroups have clinical interpretation and may contribute to the prediction of patient-specific response to therapies (treatments).
- A new method for variable selection in model-based clustering is developed, which outperformed the existing technique.
- R package was written for the proposed variable selection method in model-based clustering, and will be publicly available as an open-source.

Manuscripts from this Thesis

Five manuscripts are produced from this thesis; one is awarded at ERS conference, and already published and cited more than 30 times so far, and the others are in submission.

1. Biological clustering supports both "Dutch" and "British" hypotheses of asthma and chronic obstructive pulmonary disease. Ghebre MA, Bafadhel M, Desai D, Cohen SE, Newbold P, Rapley L, Woods J, Rugman P, Pavord ID, Newby C, Burton PR5, May RD, Brightling CE. JACI 2015[1].

2. Severe exacerbations in moderate-to-severe asthmatics are associated with decreased TH-2 and increased pro-inflammatory and TH-1 cytokine profiles in sputum and serum. Michael A Ghebre, Dhananjay Desai, Beverley Hargadon, Amisha Singapuri, Chris Newby, Joanne Woods, Laura Rapley, Suzanne Cohen, Athula Herath, Erol Gaillard, Richard May, Chris Brightling. In submission.

3. Asthma and chronic obstructive pulmonary disease overlap: biological exacerbation clusters. Michael A Ghebre, Mona Bafadhel, Dhananjay Desai, Suzanne E Cohen, Paul Newbold, Laura Rapley, Jo Woods, Paul Rugman, Chris Newby, Ian D Pavord, Richard D May, Chris E Brightling. In submission.

4. Sputum pro-inflammatory mediators are increased in Aspergillus fumigatus culture positive asthmatics. Michael A Ghebre, Dhananjay Desai, Amisha Singapuri, Joanne Woods, Laura Rapley, Suzanne Cohen, Athula Herath, Andrew J Wardlaw, Catherine H Pashley, Richard D May, Chris E Brightling. In press.

5. Variable selection in model-based clustering: a finite Gaussian mixture model. Michael A Ghebre, Chris Newby, Chris E Brightling, John Thompson. In submission.

Part I

A Statistical Framework for Modeling the Biological Heterogeneity of Asthma and COPD

Chapter 1

Introduction

1.1 Introduction

The main objective of this chapter is to provide a broad introduction of asthma and chronic obstructive pulmonary disease (COPD); in which, the diagnosis, classification, risk factors, similarities and differences between the diseases will be described. In addition, the direct effect of cytokines or in relation to cellular profiles (such as neutrophils and eosinophils) in causing or mediating the airways inflammation, and the future cytokine-based treatment of asthma and COPD subpopulation will be discussed. Furthermore, the application of cluster analysis to identify novel subgroups of asthma and COPD, and the current and future phenotypic (subgroups) treatment of both diseases will be described. This chapter ends with the proposed statistical framework how to model the biological heterogeneity (using sputum cytokines) of asthma and COPD jointly in order to identify the common and distinctive biological subgroups of the diseases.

1.1.1 Introduction to Asthma and COPD

Asthma and COPD are heterogeneous diseases [2], and among the top 10 leading chronic diseases, and representing a major global causes of death, and consuming substantial health-care resources [3]. Asthma is a disorder defined by its clinical, physiological, and pathological characteristics [4], which is associated with episodic, completely reversible airway obstruction and airway hyperresponsiveness (an excessive airway narrowing in response to a variety of stimuli) [5], which leads to recurrent episodes of wheezing, breathlessness, chest tightness, and coughing [3]. It is a serious public health problem throughout the world, affecting people of all ages, with an estimated 300 million individuals affected globally, and remains the number one

chronic disease of childhood with 12.8 million school days missed. It is the most common occupational respiratory disorder in industrialized countries [6]. Whereas COPD is characterized by tobacco-related, gradually progressive, fixed airflow obstruction that is associated with airway components inducing the loss of lung elastic recoil, resting and dynamic lung hyperinflation, and abnormalities in gas diffusion [5]. The airflow limitation in COPD is usually progressive and related to an abnormal inflammatory response of the lung to noxious particles or gases. The chronic airflow limitation characteristic is caused by a mixture of small airway disease (obstructive bronchiolitis) and parenchymal destruction (emphysema) [7]. COPD is a major cause of chronic morbidity and mortality throughout the world [8]. It is the fourth leading cause of death in the world [9] and is projected to rank 3rd in 2030 due to continued exposure to COPD risk factors, such as smoking, and the changing age structure of the world population [7].

1.1.2 Diagnosis and Classification of Asthma and COPD

Asthma and COPD often clinically diagnosis by symptoms such as episodic breathlessness, wheezing, cough, dyspnea and chest tightness [4, 7]. For example, a patient can be diagnosed for asthma or COPD based on the signs and symptoms, medical and family history and spirometry test. The physician can check the patient for smoking history, exposure to lung irritants (such as contact with smokers, chemical fumes, dust and air pollution), ongoing coughing (production of sputum during coughing), wheezing and other abnormal chest sounds. In addition, tests for lung function using spirometry; how much air the patient able to breath in and out, and how fast can breathe the air out. This involves taking a deep breath in and exhaling as fast as the patient can do through a mouthpiece connected to a spirometer. The spirometer takes two measurements: the volume of air the patient can breathe out in the first second of exhalation known as forced expiratory volume in one second (FEV_1), and the total amount of air the patient breathe out that is called forced vital capacity (FVC).

A patient may be given a reliever inhaler medicine (bronchodilator) that is used to open up the airways, and then blow air into the spirometer tube again, which is called post bronchodilator, to assess whether the medicine improves the breath-

ing. Thereafter, the pre- and post- bronchodilator FEV_1 results (before and after taking the bronchodilator) can be compared, which is known as reversibility testing. The term reversibility is usually applied to rapid improvements in post FEV_1 (after administration of bronchodilator), which is commonly used in distinguishing asthma from COPD, in which the reversibility and variability provides confirmation of asthma diagnosis [4]. In contrast, the irreversible (fixed) airflow obstruction confirmed COPD diagnosis [7].

In addition, using the spirometry test results, the severity of airflow limitation in asthma and COPD can be defined for setting treatment goals. A useful assessment of airflow limitation is the ratio of FEV_1 to forced vital capacity (FVC), which is normally greater than 0.75 to 0.8, and any values less than these suggest airflow limitation [4, 7]. Asthma classified phenotypically as mild, moderate, or severe according to Global Initiative for Asthma (GINA) guideline [10–12], which are largely determined by lung function measurements (percentage predicted FEV_1 and FEV_1/FVC ratio) [13, 14]. Whereas COPD severity is classified into four stages according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) guideline using spirometry [7]; in which, stage I: Mild; stage II: Moderate; stage III: Severe; stage IV: Very severe. However, there is an imperfect relationship between the degree of airflow obstruction and the presence of symptoms [7]. For example, a patient can be diagnosed for COPD on the bases of lung function measurement using spirometry before his/her symptoms develop.

1.1.3 Risk Factors of Asthma and COPD

Asthma and COPD are highly complex and heterogeneous diseases, which are still not completely understood their risk factors. However, a number of factors that influence a person's risk of developing asthma or COPD have been determined. These can be possibly classified as host factors (primarily genetic) and environmental factors. The lack of clear definitions of asthma and COPD present a significant problem in examining the role of different risk factors. For example, the characteristics that define asthma (e.g. hyperresponsiveness, atopy and allergic sensitization) are themselves products of complex gene-environmental interaction, and are therefore both features of asthma and risk factors for the development of the disease [4].

However, in people with asthma tobacco smoking is associated with an accelerated decline in lung function [15]. In addition, environmental factors such as domestic mites, furred animals (such as dogs, cats and mice), cockroach allergen, fungi, molds, yeasts, pollens, infections (predominantly viral), occupational sensitizers, outdoor/indoor air pollution and diet play considerable role in causing asthma [4]. Whereas COPD prevalence, morbidity, and mortality vary across countries and across different subgroups within a country; however, these are directly related to the prevalence of tobacco smoking [7]. Although cigarette smoking is a well-established risk factor of COPD [16], others such as genetic factors, longstanding asthma, indoor and outdoor air pollution, passive smoking exposure, biomass smoke, occupational exposures, diet, and tuberculosis are identified as possible independent risk factors of COPD [16]. There is also evidence that the risk of developing COPD is inversely related to socioeconomic status [17].

1.1.4 Similarities and Differences between Asthma and COPD

Although the clinical symptoms of both diseases are caused by airway narrowing as a result of inflammation in the airways [7], it is still not fully understood their sharing and distinctive characteristics due to the complex and heterogeneous nature of the diseases [2]. Over the last several decades, there has been a considerable discussion in respiratory literature, largely after the Dutch and British hypotheses were reported. The Dutch hypothesis suggested that all obstructive diseases (including asthma and COPD) are manifestations of the same basic disease process; whereas the British hypothesis suggested that asthma and COPD are two distinct entities generated by different mechanisms [18].

Recently, PJ Barnes suggested that asthma and COPD have marked differences in terms of cellular mechanisms, inflammatory mediators, and response to therapy, but they also share a number of characteristics in which some patients with COPD also had characteristics of asthma [19, 20]. Similarly, Welte & Groneberg reported that these two diseases are distinctive along all stages of severity with some overlap [21]. In addition, Bianchi et al, suggested that individuals with COPD may have features of asthma such as a mixed inflammatory pattern with increased eosinophils [22]. Furthermore, Kesten et al, reported that COPD can coexist in individuals

with asthma (especially with severe asthma) who are exposed to noxious agents, particularly cigarette smokers, and may develop a fixed airflow limitation and a mixture of "asthma- like" and "COPD-like" inflammation [23, 24]. There is also epidemiological evidence that long-standing asthma on its own can lead to fixed airflow limitation [25].

Over the last five years a number of studies [2, 26–34] have performed a comprehensive review regarding asthma and COPD overlap, and fairly concluded that there is overlap between the two diseases, but they did not categorically stated the degree of overlap at different levels of the diseases.

1.1.5 Profiling of Bacterial Communities in Asthma and COPD

The bacterial connection with asthma and COPD is well-documented. However, the role of bacterial infection in the pathogenesis of asthma and COPD, and how it should be treated has been an ongoing source of debate. Bacterial infections are involved in almost 50% of COPD exacerbations. However, only a few of pathogens have been consistently identified in the airways which were mainly using culture-based approach, and the bacterial microbiota in acute exacerbations remains largely unknown [35]. Previously, it has been suggested that infections (viral and bacterial) may contribute to the pathogenesis and progression of COPD [36], and bacteria may induce inflammation at both stable and exacerbation states [37]. Some COPD studies also showed the relationship of COPD pathogenesis and exacerbations with bacteria colonisation and infection [38], and an association between bacterial colonization and airway inflammation [39], and airway bacteria load and decline in FEV_1 [40]. Whereas in asthmatic studies, it has been suggested that bacterial organisms may increase airway hyperresponsiveness and inflammation [41], and has been described the role of bacteria colonization in perpetuating inflammation in the lower airways [42–44]. In addition, asthmatic patients with neutrophilic inflammation are commonly culture-positive for Haemophilus influenza [42, 43], which may suggest the potential role of bacteria presence (especially H. influenza) in the lower airway in the continuation of neutrophilic airway inflammation [44].

The accurate diagnosis and treatment of bacterial infection in individual patient

remains a major challenge. Antibiotics have been used as standard management for the treatment of exacerbations of COPD patients, but their impact remain unreliable [45]. The trials that have assessed the effect of antibiotics in COPD subjects generally are not good quality and were not well controlled. A major challenge remains how to define the potential role of bacteria in the inflammatory process and how best can optimize the use of antibiotics without overutilize in order to avoid drug resistance. Alternative strategies to treat infection in COPD remain very limited; however, recent trials of the long-term use of macrolides have shown promising results [46], and found significant reduction in the rate of exacerbations.

Since the aforementioned data were predominantly derived from culture-based approach, the presence of other important organisms involved in asthma or COPD that were not easily cultured were not explored fully. Recent technological advances in diagnostic techniques, particularly the use of 16S sequencing has demonstrated that there are a large range of bacteria present in the lower respiratory tract, and the secrets of the human microbiota are beginning to be unravelled, and reveal the existence of a complex and diverse array of bacterial communities. The 16S rDNA sequence approach do not rely on growing organisms in pure culture (it is a culture-independent technique), in which bacterial community profiles were generated using the high-throughput sequencing that makes the detailed assessment of airway bacterial colonisation possible. The microbiome is defined as the total collection of microbiota that resides within humans or on their skin surface [47]. Thus, the lung microbiome is the complete collection of microbiota living in the airways and parenchymal tissues [47].

Recently, asthma and COPD studies started to perform culture-independent microbial community profiling to characterise the lower airway microbiome which generated from 16S rDNA based sequencing with the aim to assess a wide range of microbial communities abundance at different taxa levels including bacteria phyla (such as Firmicutes, Actinobacteria and Bacteroidetes and Proteobacteria) and genera (such as Haemophilus, Streptococcus and Moraxella). In asthmatic study, it has been found that Haemophilus appeared abundant in a younger atopic men subgroup who have elevated level of neutrophils [44]. Despite the abundance, the functional role of Haemophilus in causing/mediating the neutrophilic inflammation and po-

tential implications for pathogenesis of the disease is still unknown. In addition, the prevalence of other bacterial phyla and/or genera in other airway inflammations such as eosinophilic are poorly understood although the prevalence of Firmicutes, Actinobacteria, Bacteroidetes and Proteobacteria appeared to increase significantly in non-neutrophlic asthmatics subgroup [44].

So far, no single study has compared the patterns of microbiome communities (composition), at either phylum or genus level, across asthma and COPD neither at stable nor at exacerbation states. Thus, the patterns of microbiome communities profiling (which not represented by counting the bacterial loads) across both diseases are largely unknown. Investigating the changes in the relative abundance of members of the microbial communities across asthma and COPD with respect to the frequency and severity of exacerbation due to cellular inflammation would be a step forward to understand the role of bacteria in the airway diseases, and may unfold new insights and approaches to the pathogenesis and treatment of lung infection in both diseases.

1.1.6 Role of Cytokines in Asthma and COPD: In Relation to Cellular Profiles

Cytokines are extracellular signaling proteins expressed by different cell types involved in cell-to-cell interactions [48]. Asthma and COPD are characterized by chronic inflammation and remodeling of the airways [49, 50], and cytokines play a significant role in orchestrating the chronic inflammation and structural changes of the respiratory tract in both diseases [48]. Recently, cytokines have become important targets for the development of new therapeutic strategies for both asthma and COPD [20].

For more than a decade, a considerable effort has been made to define asthma and COPD at an inflammatory cellular level with the aim to understand the potential mechanisms in order to improve the diseases management and treatment. The role of cellular profiles such as eosinophils and neutrophils in the airways inflammation are well documented, and it has been recognized that the airways inflammation in asthma is characterized by an eosinophilic and in COPD predominantly by neutrophilic inflammation [19, 51–53].

Eosinophils are white blood cells that consist about 2 to 4% of the total leuk-

ocytes (white blood cells), and they increase in response to allergies and parasitic infections, and are infrequent in the blood, but many in the mucous membranes of the respiratory tracts [54]. For example, in most asthma phenotypes, eosinophils levels incremented in the tissues, blood and bone marrow, and are positively correlate with the disease severity [55]. Whereas neutrophils are the most abundant white blood cells, consists of about 60-70% of the circulating leukocytes [56], and increased level of neutrophils were positively associated with the severity of COPD [57]. Both eosinophils and neutrophils are essential in healthy lungs and important component of innate immunity that protect individuals against infection. However, elevated level of eosinophils in the airways may cause exacerbations and irreversible damage to the airways [58]. Similarly, for example in COPD, increased level of neutrophils in the inflammation sites could be harmful and may play a key role in the destructive processes that could be responsible for potential damage of healthy tissues [59].

Several cytokines (including IL-5, IL-6 and TNFα) [20, 60–62] and cellular profiles (such as eosinophils and neutrophils cell-counts) [19, 51–53] were associated with asthma or/and COPD pathogenesis, but the direct role of the cytokines or in relation to the cellular profiles in causing or mediating the inflammation were not fully understood. However, T_H2 derived cytokines (such as IL-4, IL-5 and IL-13) were identified as critical inflammatory mediators in orchestrating the eosinophilic inflammation [60, 63]. In addition, increased level of key proinflammatory cytokines including IL-1β and tumor necrosis factor-a (TNFα) prolong eosinophil survival in asthmatic airways [7]. Furthermore, cytokines such as TNFα, IL-1β and IL-6 amplify the inflammatory process (involving neutrophils) and may contribute to some of the systemic effects of COPD [8]. The identified cytokines which are expected to involve in asthma and COPD, and their networks are displayed in figures 1.1 and 1.2, respectively.

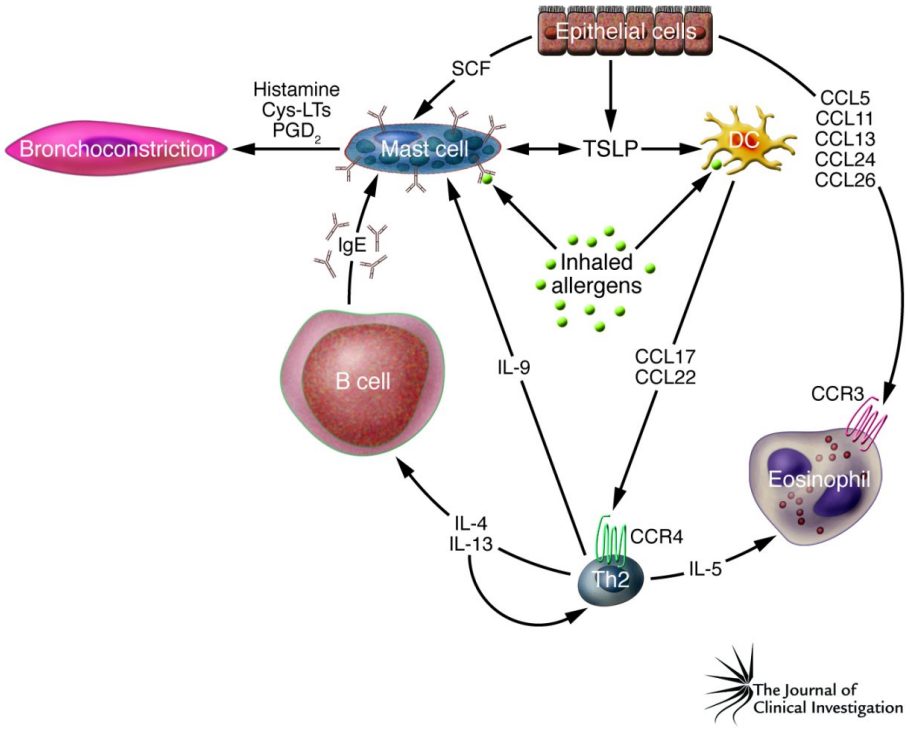


Figure 1.1: Cytokines involved in Asthma

(The figure is reproduced from Barnes PJ, J. Clin. Invest. 118:3546–3556 (2008). [20])

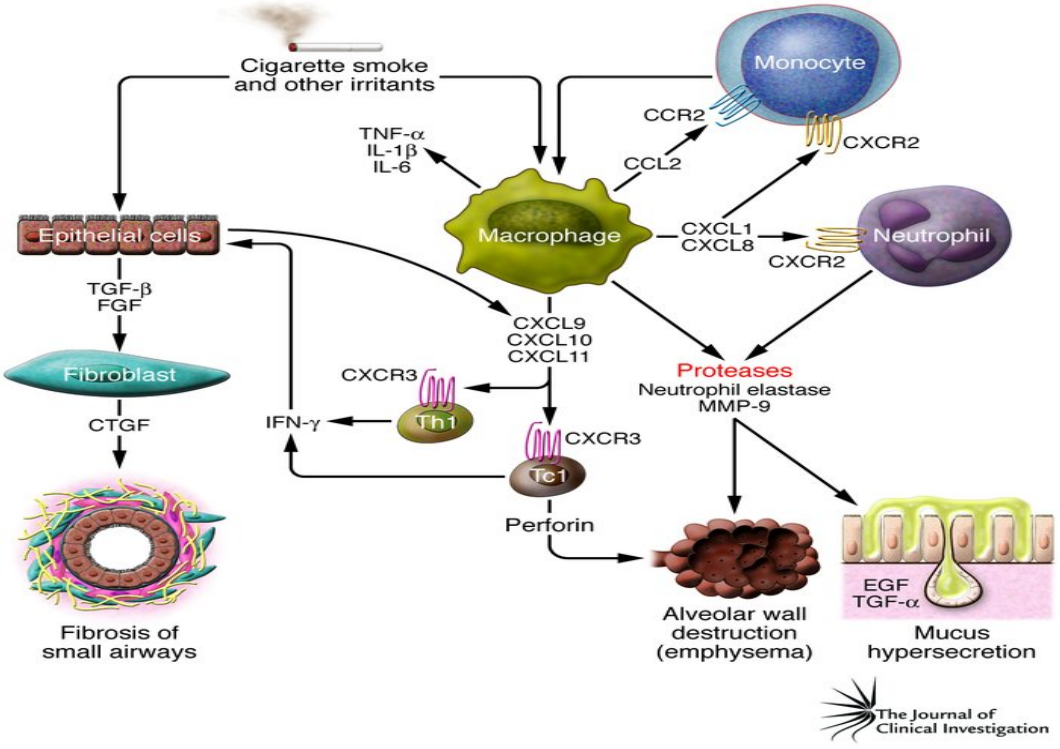


Figure 1.2: Cytokines involved in COPD

(The figure is reproduced from Barnes PJ, J. Clin. Invest. 118:3546–3556 (2008). [20])

It is well established that asthma and COPD are heterogeneous diseases with respect to clinical characteristics, cellular sources of inflammation, and response to common therapies. However, understanding the entire heterogeneity remains elusive, and the differences in cytokine-driven inflammation (such as eosinophilic and neutrophilic) may underlie some of the heterogeneity. Thus, cytokine profiles may serve as main targets for development of novel anti-inflammatory drugs to reduce the inflammation and subsequent exacerbations. However, the patterns of the mediators with respect to the cellular profiles and other outcomes (such as diseases lung function severity and frequency of exacerbation) across different population subgroups should have to be investigated for effective treatment.

1.1.7 Future Cytokine-based Treatment Targeting Asthma and COPD Subgroups

Current guidelines outline the diagnostic and management strategies for both asthma and COPD, and the standard treatment of subjects is predominantly a combination of inhaled corticosteroids and long-acting beta2-agonists. However, the symptoms of many patients remain poorly controlled as the treatment approach and management strategies do not consider the heterogeneity of the diseases population. The presence of eosinophilic and neutrophilic inflammation, and reversibility of lung function (based on spirometry measurements) were earlier viewed as distinguishing features between asthma and COPD. However, this paradigm has been changed recently due to the recognition of the heterogeneity of the diseases. For example, asthmatics and COPD subjects appeared to share common physiologic abnormalities of airflow limitation (obstruction on lung function); symptoms such as shortness of breath, chest tightness, wheezing and coughing. In addition, several studies showed that fraction of asthmatics do not have elevated level of sputum eosinophils, but those subjects (with non-eosinophilic inflammation) have increased features of neutrophilic inflammation [64]. Furthermore, eosinophilic inflammation (differential sputum eosinophil cell count $> 3\%$) was observed in up to 40% of COPD patients [65].

Currently, there are very limited therapies for both asthma and COPD. Thus the development of novel treatment strategies is an urgent need, but it requires a deeper understanding of the underlying inflammatory processes associated with

the diseases pathogenesis that accounts the diseases heterogeneity. The current treatment (therapeutic trials) has largely relied on the improvement of FEV_1 as a main outcome (the primary measure of disease severity), however FEV_1 remains a poor surrogate marker for disease activity [66–68]. Thus, effective alternative surrogate biomarkers (beyond FEV_1) that account the population heterogeneity are needed to predict diseases outcomes such as inflammation, and frequency and severity of exacerbation.

In recent years, increased understanding of the diseases heterogeneity in terms of eosinophils and neutrophils cellular profiles has led to improved knowledge of the pathogenesis of asthma and COPD and allowed new approach of treatments to be investigated, and have already given clinicians an effective framework to use the available treatments. For example, those asthmatics with elevated level of sputum eosinophilia tend to have improved response to inhaled and systemic corticosteroid treatment [69]. Similarly, in COPD subjects, eosinophilic inflammation is associated with favorable response to corticosteroid therapy [70], and reduction in severe exacerbation rates [71]. In addition, eosinophilia (differential sputum eosinophils cell-count $> 3\%$) was identified as a key mediator in differentiating the use of new asthma targeted treatment known as mepolizumab (a humanized monoclonal blocking antibody against IL-5) [58, 72, 73], and becomes effective in reducing the inflammation in the airways for asthmatics who have elevated level of eosinophils. This may provide a similar benefit for COPD patients, who have similar eosinophils pattern, which is currently under investigation [74].

An early placebo-controlled trial study with mepolizumab (anti-IL-5) was not effective [75], but the outcome improved with patients' classification based on sputum eosinophils elevation (defined by sputum eosinophil percentage $> 3\%$). Particularly two trials showed the benefits of mepolizumab in subjects with severe asthma who were selected based on elevated sputum eosinophilia [58, 72]. In the first trial, it has been found a reduction in the frequency of exacerbations when the drug was given to subjects with refractory eosinophilic asthma (despite they were on high-dose of corticosteriod treatment) [72]. The second study was also performed in asthmatic subjects with sputum eosinophilia and airway symptoms (despite continued treatment with prednisone) and found a reduction in the number of blood and sputum

eosinophils [58]. However, subjects who had taken the new anti-IL-5 drug did not show consistent improvement on their lung function measurements (i.e. FEV_1). This pattern suggests that we need to consider other diseases outcomes that could predict the inflammation and frequency of exacerbation. For instance, FEV_1 appeared as a poor marker to predict the exacerbations caused due to cellular profile inflammation that is mediated by T_H2 high cytokines [72].

These positive studies highlight the need to classify asthma and COPD population into subgroup for better understanding of the diseases pathology and heterogeneity, and subsequently develop a new therapeutic or use existing standard treatments targeting specific subpopulation. Whether other non-T_H2 (such as T_H1 or pro-inflammatory) cytokine pathways underlie airway inflammation in specific subsets of asthma or/and COPD patients is an unresolved question. Some of the cytokine effects on various airway components with a T_H1/T_H2 imbalance in mild and severe diseases were identified and summarized in figure 1.3 on page 19. However, the potential mechanisms of these networks are not fully understood.

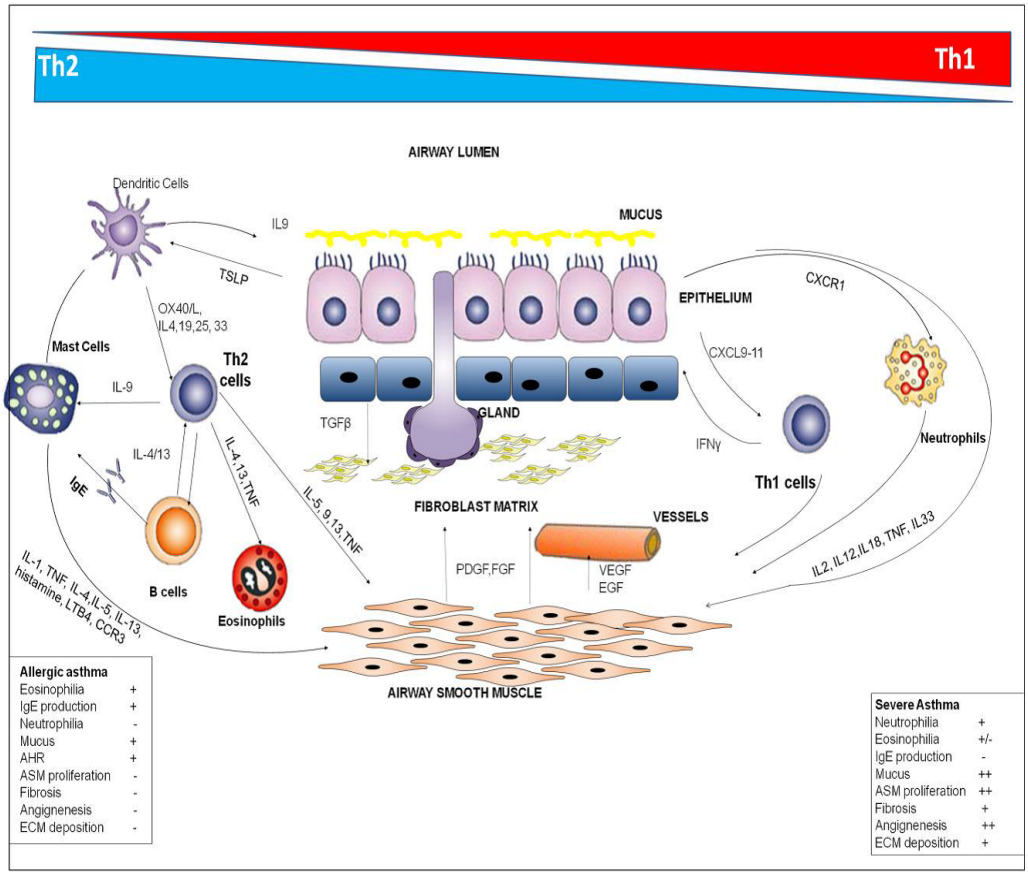


Figure 1.3: Cytokine effects on various airway components with a T_H1/T_H2 imbalance in mild and severe disease.

(The figure is adapted from ATS/ERS guidelines 2013 [76])

It has been observed that neutrophilia airways inflammations (sputum neutrophil differential cell counts above 61% [77]) was seen most often in subjects with severe asthma and COPD, who are more likely on high doses of corticosteroids or prednisone treatment. For example, neutrophilia has been observed in asthmatics with acute and chronic infection [78], in subsets of severe asthmatics [79], and during acute asthma exacerbations [80], and was associated with reduced lung function independent of eosinophils [81, 82]. Mixed neutrophilia and eosinophilia inflammation was also reported in refractory asthma study, and found that these subjects had lowest lung function, highest frequency of daily wheeze, and highest health care utilization [83]. Thus, further classification of subjects with asthma and COPD according to cellular profiles may lead to four distinctive subgroups, which are: (I) pure eosinophilic (eosinophil $> 3\%$ and neutrophil $< 61\%$), (II) pure neutrophilic (eosinophil $< 3\%$ and neutrophil $> 61\%$); (IV) mixed granulocytic (eosinophil $> 3\%$ and

neutrophil > 61%); (IV) and paucigranulocytic (eosinophil < 3% and neutrophil < 61%). The importance of this classification approach may provide an understanding of the disease pathology in each group and facilitate the development of new therapeutic approach targeting each subgroup. In addition, it may help to assess whether the strategies for patients' selection based on this classification will improve the outcomes of the clinical trials, or adjustment of the standard treatment based on the threshold of a particular predictive biomarker in each group. For instance, understanding or predicting the above possible subgroups based on the patterns of cytokine profiles may help to develop cytokine-based therapeutic targeting each group similar to the mepolizumab (T_H2 cytokines-based anti-inflammatory drug) that appeared to be effective in reducing the eosinophilic inflammation and subsequent exacerbation. The effectiveness of mepolizumab was clearly related to how precise the patients were classified in terms of sputum eosinophilia.

There is a major advantage of using modern mathematical/statistical techniques to characterize asthma and COPD biological heterogeneity based on cytokine profiles. Creating subgroups based on the activity (or lack thereof) of specific cytokine profiles may identify mechanisms underlying each group, a new potential biomarker that could predict treatment response or/and useful for patients' selection in clinical trial (to maximize the chances of success). For example, the pathophysiologic connection of cytokine profiles with cellular profiles and/or other diseases outcomes in each group may aid in the development of new cytokine-targeted therapies and may identify biomarker that have the potential to inform clinical trial investigators as to which subjects may respond to the therapies under study and why. In conventional clinical trials, the efficacy of the drug under study is based on the average response compared to placebo or a standard treatment (control group); the problem with this approach is that the presence of a difference in average response can be driven by a few outliers that may lead to the acceptance of the new treatment that only helps a specific subpopulation. On the other hand, the lack of a significant difference in the average response (primary outcome) across treatment and control groups may lead to the rejection of the new drug even though it may be beneficial for a particular subgroup as it has been observed in early study of mepolizumab [75]. In addition, the identified biomarkers may also provide guidance to physicians to reduce or in-

crease standard existing treatment as it might be cheaper compared to the new drug (cost-effective). For example, knowledge of the state (threshold) of airway inflammation would allow for the adjustment of anti-inflammatory medication such as inhaled corticosteroids, with the possibility of preventing subsequent exacerbations. For example, three randomized controlled trials have studied the utility of following induced sputum eosinophil counts in moderate-to-severe asthmatics to adjust the dose of inhaled corticosteroids, and found a reduction in the frequency and severity of asthma exacerbations when the inhaled corticosteroid dose was adjusted based on induced sputum eosinophil counts compared to the standard-of-care method of dose adjustment by symptoms, lung function, or rescue medication use [84–86]. The association between sputum eosinophils and a favorable response to inhaled corticosteroids suggests a potential role for this signature as a biomarker to guide the decision to start, continue, or change the dose of inhaled corticosteroids.

The anti-IL-5 drug emerged as effective for reducing eosinophilic inflammation that is mediated by T_H2-high cytokine. However, whether other inflammations such as pure neutrophilic or mixed (eosinophilic and neutrophilic) are mediated by non-T_H2 mediators such as high in T_H1 or pro-inflammatory is unknown and unstudied. Therefore, the possibility of a larger degree of airway inflammation needs investigation. It is possible that other subset of patients with T_H2-low may have the disease on the basis of T_H1 or pro-inflammatory mediators in which these cytokine pathways underlie the airway inflammation. For example, understanding and distinguishing the non-T_H2 cytokine pathways underlie these and other airway inflammation in T_H2-low subpopulation may help to develop non-T_H2 cytokine-based therapeutic approach targets these pathways in specific subsets of patients, and identify new biomarker that play a fundamental role in the clinical manifestations of a subset of patients. Therefore, the mechanisms underlying these subgroups can be uncovered with improved stratification of the population on the bases of a panel of cytokine profiles patterns using appropriate statistical techniques (such as cluster analysis). Cluster analysis has the potential to dissect the population heterogeneity based on the patterns of input variables (cytokines) and identifies distinctive subgroups that may have similar disease pathology (underlying pathway).

1.2 Application of Cluster Analysis to Model Asthma and COPD Heterogeneity

Cluster analysis is an unsupervised statistical technique that uses to uncover unknown/hidden structures among heterogeneous individuals (subjects) by seeking partition of the subjects into distinct subgroups based on the input variables. The observations within each cluster are more similar (homogeneous) to each other than between clusters. It is a useful technique to understand the variation and explore hidden structures of complex data, and generating hypothesis for supervised studies such as linear regression and discriminant analyses.

The application of cluster analysis to model population heterogeneity in medical and social science studies is widespread in order to identify novel subgroups/clusters with specific characteristics. Recently, several studies provide a view of asthma and COPD heterogeneity through multivariate clustering approach. For example, Haldar and colleagues [87] performed k-means cluster analysis on three asthmatic populations (predominantly mild-to-moderate asthmatics from primary care, a refractory asthmatic from secondary care, and predominantly refractory asthmatics from randomized trial in standard care) and identified five distinctive clusters. Which were best represented by varying symptom control with the degree of eosinophilic inflammation (sputum differential cell count), in which the first two clusters (early-onset, atopic and obese non-eosinophilic) were common to both primary care and the refractory asthmatics. The second two clusters were characterized by marked discordance between symptom expression and eosinophilic airway inflammations (early-onset symptom predominant and late-onset inflammation predominant) were unique to refractory asthmatics. The fifth cluster was labeled benign asthma as the subjects within this group had little evidence of active disease.

Likewise, Moore et al. [14] also identified five asthmatic clusters using hierarchical cluster analysis using only spirometric measurements and clinical characteristics; in which subjects in cluster 1 have early onset atopic asthma with normal lung function (who were treated with two or less controller medications and minimal health care utilization). Cluster 2 consists of subjects with early-onset atopic asthma and preserved lung function but increased medication requirements and health care util-

ization. Cluster 3 characterized predominantly by older obese women with late-onset non-atopic asthma (who showed moderate reductions in FEV_1 and frequent oral corticosteroid). Subjects in clusters 4 and 5 have severe airflow obstruction with bronchodilator responsiveness, but they are different with respect to their ability to achieve normal lung function, age of disease onset, use of oral corticosteroids and atopic status. Siroux et al. [88] also identified four asthmatic clusters using latent class analysis (clustering approach). The subgroups were discriminated based on the quality of life, and blood neutrophil counts. The first two clusters were characterized as active treated allergic childhood-onset asthma and active treated adult-onset asthma. The last two subgroups consists of subjects with inactive or mild untreated asthma (who differed by atopy status and age of asthma onset). Similarly, Just and colleagues (2012) [89] applied cluster analysis to identify subgroups of childhood asthma, and identified three distinctive clusters. Cluster 1 consists subjects with severe exacerbations and multiple allergies; cluster 2 represents subjects with severe asthma and bronchial obstruction; and cluster 3 consists subjects with mild asthma.

Cluster analysis was also applied to model COPD heterogeneity to identify distinctive subgroups. For example, Bafadhel and colleagues [90] applied the combination of factor and cluster analyses to identify exacerbations biological subgroups in COPD. They identified four distinctive clusters, in which cluster 1 has high proportion of bacterial colonization; cluster 2 consists of subjects with high proportion of viral; cluster 3 comprises of subjects with elevated level of sputum eosinophils; and cluster 4 consists of subjects with limited changes in the inflammatory profile. These clusters were labeled as bacterial, viral, eosinophilic-predominant, and pauci-inflammatory subgroups, respectively. Similarly, Burgel et al. [91] applied the combination of principal component and cluster analyses to identify COPD subgroups using multiple clinical variables, and identified four clinical clusters beyond the GOLD standard classification; in which subjects with varying airflow obstruction (FEV_1) were assigned to different subgroups and had noticeable differences in age, comorbidities, symptoms and predicted mortality.

Although cluster analysis has been applied separately in asthma and COPD studies, to my knowledge (particularity at biological or cytokines level) this technique was not yet implemented to the combination of asthma and COPD studies

in order to identify the common and distinctive subgroups of the diseases. So far, the diseases similarities and differences were predominantly identified by comparing the subjects characteristics across asthma and COPD at disease level as defined in the clinic. This simple separation approach may not reflect the phenotypic heterogeneity of the diseases, and the crude comparison of the characteristics at higher level may not capture the hidden subpopulation of the diseases. In clinic, a subject is diagnosed for asthma or COPD based on the current guidelines that includes patient perception of symptoms (that vary hugely from one patient to another and from time to time within a patient) by incorporates the physicians perception from their monitoring of the individuals condition. This approach is subject to considerable bias and variability and shortfall of providing insight the multidimensional characteristics and heterogeneity of the diseases population.

However, cluster analysis can uncover the common and distinctive new subgroups of both asthma and COPD based on the entire patterns of measured variables (characteristics); and comparisons can be implemented across the subgroups rather than the diseases. In cluster analysis, the disease status (whether a subject was diagnosed as asthmatic or COPD by a physician) is avoided whilst modeling the heterogeneity of the population.

1.3 Proposed Approach to Model the Biological Heterogeneity of Asthma and COPD

The characterization of asthma and COPD subgroups based on cytokine profiles and the recognition that these subgroups are associated with significant variability in responses to the established and/or emerging therapies has motivated this study. The need to refocus efforts to define the similarities and differences of asthma and COPD in terms of cytokine profiles [92] is underscored by the development of highly specific anti-inflammatory therapies because response is more likely to be subgroup related instead of disease-specific [93]. This is perhaps best characterized by anti–IL-5 approaches, which have demonstrated clinical responses related to underlying eosinophilic lung inflammation in asthma which is mediated by T_H2 high cytokines [58, 72] and similar strategies are currently being tested in COPD [74].

Whether other non-T_H2 cytokine pathways underlie neutrophilic or other airway inflammations in specific subsets of asthma or/and COPD patients is an unresolved question. To enable these and further analogous developments, there is an urgent need to define the airway cytokine profiles in both diseases. Modeling (clustering) the biological heterogeneity of asthma and COPD jointly using sputum cytokines may identify subpopulation who could benefit from targeted approaches that enhance the efficacy of the current and/or future cytokine-based treatments. Although the existing guidelines are valuable, a better appreciation of diseases heterogeneity and application of the emerging cytokine-based treatment targeting specific subpopulation will be important as currently there are very limited therapies for both diseases.

In this study, we will apply cluster analysis on the panel of sputum cytokine profiles to identify asthma and COPD common and distinctive biological subgroups (distinguishable based on the differences in the composition of the cytokine profiles) without taking into account the demographic, lung function and cellular markers and disease status. However, the identified subgroups' clinical relevance/utility will be assessed based on the clinical outcomes such as cellular inflammations (eosinophilic or/and neutrophilic), frequency of exacerbation, bacterial colorizations or/and disease severity (lung function measurements).

However, clustering using the actual cytokines measurements may not be straightforward as these mediators have strong correlations (hidden structures) that are currently ignored by the standard clustering techniques. To my knowledge, there is no previous robust statistical framework (which was applied or proposed to model the biological heterogeneity of asthma and COPD using cytokine profiles), except the machine learning approach that was used in asthmatic study [94] and a 2-mode graphical approach with Kamada-Kawai algorithm was applied to identify the patterns of the mediators and subjects' subgroups. However, this study did not report the summary statistics of the mediators or other clinical characteristics across the identified subgroups, except the graphical visualization of the cytokines and clusters. In addition, the algorithm they have applied (i.e. Kamada-Kawai algorithm) has serious drawbacks as many links crossings appear producing a lot of nodes overlap, and making the reading of the graph harder in which edges can go backwards to

the center of the map and placing close two nodes linked by long path that this can give a false impression of closeness because of the spatial distribution of the nodes. Thus, this approach may not be an appropriate for identifying and visualization of the clusters using cytokine profiles, and their graph should have to be interpreted cautiously.

Although the application of cluster analysis in respiratory research has been widely used to discover new subgroups, the corresponding statistical issues in planning to use the right inputs variables (which account the internal patterns/correlations of the variables) in clustering algorithm receive rather less emphasis. This part of the study will emphasis in identifying the right input variables (which represent the measured sputum cytokines) into a clustering algorithm to model the biological heterogeneity of asthma and COPD, and subsequently to identify the common and distinctive subgroups of the diseases. To identify an appropriate approach, an artificial data will be simulated, which has similar internal patterns as the cytokines, but with known class membership. The artificial variables will be represented by several variables, such as latent variables or highest loading variables (deriving from factor analysis), and will be independently used as input into a clustering algorithm. The statistical issues/bias of each approach in identifying the optimal clusters will be discussed. Then the method that performs best in identifying the optimal clusters in the artificial data will be applied to asthma and COPD cytokines study.

This study expected to provide a general guideline material for researchers that may suggest which statistical approach could be appropriate to model the population heterogeneity using correlated variables (such as sputum cytokines). In addition, it will show the consequences of hidden patterns/structures of the input variables in concealing the optimal clusters. Furthermore, this project may enhance our understanding of asthma and COPD heterogeneity, with respect to sputum mediators at biological level, and may provide new insights that acknowledge the overlap and highlights the differences between the two diseases beyond the current guidelines.

Hypothesis

The diseases dimension is so complex and heterogeneous but the available treatment is very limited. Therefore, I hypothesis that the mechanisms underlying asthma and COPD can be uncovered with improved stratification by identifying specific subgroups with distinct characteristics. Statistical techniques can identify these subgroups and characteristics, and hence will enable a personalised or stratified medicine approach to therapy and direct future mechanistic studies.

1.4 Objectives

- To identify a robust statistical framework to model the biological heterogeneity and uncover hidden patterns in the data of asthma and COPD population.
- To investigate the common and distinctive biological subgroups of asthma and COPD at stable state.
- To develop a classifier model for further subject assignments to the identified biological subgroups.
- To investigate the common and distinctive biological subgroups of asthma and COPD at exacerbation state.
- To assess the patterns of microbiome communities across the exacerbated asthmatic and COPD subjects, and across the identified exacerbation biological subgroups.

Main Structure of the First Part of the Thesis

The goal of this part of the thesis is to identify a robust statistical approach to model the biological heterogeneity of asthma and COPD, using sputum cytokines. Thereafter to identify the common and distinctive subgroups of the diseases at both stable and exacerbation states. The subsequent chapters, in this part of the thesis, cover the following:

1. In chapter 2, an exploratory analysis was preformed in order to understand the similarities and differences between asthma and COPD at different levels, and the underlying structures/patterns of the inflammatory mediators (sputum cytokines) were investigated.

2. In chapter 3, artificial data (with known class membership) that have similar internal patterns/structures as the cytokines were simulated. Several sets of variables that represent the artificial variables were identified, and were independently used as input variables into a clustering algorithm. The performance of each method was assessed.

3. In chapter 4, the method which performed best in the simulation study was applied to identify common and distinctive biological subgroups of asthma and COPD at stable state.

4. In chapter 5, the stable clusters (which were identified in chapter 4) were validated on independent asthma and COPD studies using two approaches.

5. In chapter 6, the common and distinctive exacerbation subgroups of asthma and COPD were identified using a robust statistical technique, and the patterns of microbiome communities were assessed across the identified subgroups.

Chapter 2

Asthma and COPD Heterogeneity at Stable State

2.1 Objectives

The objective of this chapter is to build a background understanding of asthma and COPD characteristics at stable state using descriptive and graphical techniques at different levels (clinical and biological). This may deepen our understanding of both diseases, and may also facilitate the ability to identify the right statistical technique in order to model (cluster) the biological heterogeneity of the diseases in the subsequent chapters.

2.2 Introduction

Asthma and COPD are complex and heterogeneous diseases. Thus far, our understanding on these diseases is limited, which partially could be due to the limitation of an application of appropriate statistical techniques to model the heterogeneity and uncover the hidden patterns or/and lack of relevant datasets. In this study, subjects from prospective asthma and COPD were participated with comprehensive demographic, clinical and biological characteristics.

In this chapter, asthma and COPD study population and their inclusion and exclusion criteria, and the methods that were applied for measuring the characteristics in laboratory or clinic will be briefed. The established cutoff of the characteristics will be described, and the similarities and differences between the two diseases at different levels (such as at demographic, clinical and biological) will be assessed. The internal patterns of the sputum mediators will also be investigated using graphical techniques.

2.3 Methods

2.3.1 Asthma and COPD Study Population

Subjects at stable visit (eight weeks free from an exacerbation) with moderate to severe asthma and COPD were included. They were recruited from the general respiratory and severe asthma clinic at the Glenfield Hospital, Leicester, UK, to enter two independent prospective observational biomarker studies [1, 90]. Asthma was defined according to Global Initiative for Asthma (GINA) guidelines or based on a physician's diagnosis and severity stratified according to the GINA treatment steps [4]. COPD patients, with a physician diagnosis, were defined according to GOLD guidelines [7].

In brief, subjects with history of severe asthma according to GINA guidelines and aged 18 were included in the asthmatic study, but subject who have evidence of non-asthma respiratory diseases such as COPD were excluded from the study. Whereas COPD subjects aged above 40 years, and their post 400mcg salbutamol bronchodilator FEV_1/FVC ratio < 0.7 who had at least had one exacerbation in the previous year which requires corticosteroids and/or antibiotic therapy, or hospitalized for an exacerbation were included in the COPD study. COPD subjects who unable to produce sputum following the induced sputum procedure and/or current or previous history of asthma were excluded from the study. However, the presence of co-morbidities, reported atopy to common aeroallergens, or significant reversibility on lung function testing was not an exclusion criteria in the COPD subjects. All patients recruited provided written informed consent and could voluntarily withdraw from the study at any time. The study was approved by the local (Leicestershire, Northamptonshire and Rutland) research ethics committee. Details of the study recruitment and examination process have been described elsewhere [1, 90].

2.3.2 Measurements

Information such as demographics, lung-function tests was collected. This included a medical history, smoking history and a detailed prescription history; pre- and post-bronchodilator FEV_1, FVC and symptoms recorded using the visual analogue

scale (VAS) for the domains of cough, dyspnea. In addition, spontaneous or induced sputum was collected for sputum total cell and differential counts (neutrophil and eosinophil), cytokine profiling and 16S rDNA based microbiome communities. In addition, blood were collected for full blood count and differential cell count. Here, the protocol for sputum induction, the collection and processing of sputum and blood, and process of extraction the microbiome communities will be described briefly.

Sputum and Blood Collection and Processing

Sputum Induction Protocol

Spontaneous or induced sputum was collected from subjects during visits throughout the study. Both methods of sample collection have been shown to be similar for the differential cell counts [95]. The subjects who were unable to spontaneously produce sputum, the following sputum induction protocol was carried out.

1. Guidance on position: sit upright during the nebulisation procedure and lean forward during expectoration.

2. Guidance on effective expectoration: instructions for coughing and moving sputum successfully into specimen container

3. Guidance on contamination reduction: instructions to blow nose and to rinse mouth prior to expectoration.

The procedure requires all subjects to have FEV_1 measured before and after pre- treatment with 400μg inhaled salbutamol to minimize bronchoconstriction. Nebulised saline (5mL at 3, 4, and 5%) was given in sequence via an ultra-sonic nebuliser (Ul- traNeb, DeVilbiss, Sunrise Medical, USA) for 5 minutes. After each inhalation, subjects were asked to blow their nose and rinse their mouth prior to coughing and expectoration of sputum. FEV_1 was measured after each inhalation to assess for bronchoconstriction and to assess safety for procedure continuation. The process was terminated if there was more than 20% drop in FEV_1, significant symptoms or successful sputum expectoration. The sputum induction protocol used

is depicted in figure 2.1 on page 32. All sputum samples were processed within two hours of collection in a Class II biological safety cabinet.

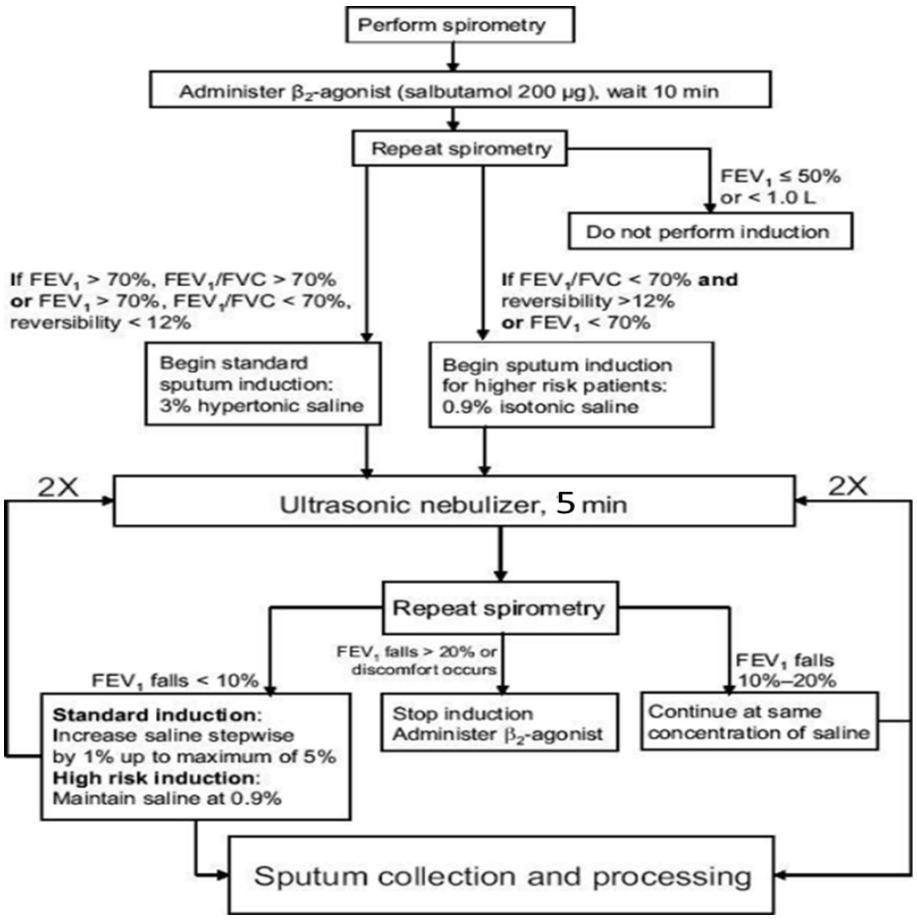


Figure 2.1: Sputum induction protocol

Sputum Collection and Processing

The collected sputum sample was emptied into a petri dish and placed on a dark background to aid visualization of sputum plugs. Sputum plugs were selected from the saliva and were then gathered into a large condensed mass in small circular movements. Sputum plugs were then removed for analysis of bacteria. The remainder of the selected sputum was weighed and incubated with 8 times the volume/weight of Dulbecco's phosphate buffered saline (D-PBS) (Sigma, Poole, Dorset). The sputum sample was dispersed by gentle aspiration into a Pasteur pipette and placed onto a bench rocker for 15 minutes on ice and then centrifuged at 790g for 10 minutes at 4^{o}C. This was followed by removal of 4 times the volume/- weight of this D-PBS supernatant with storage in 300μL aliquots at -80 ^{o}C for further mediator (cytokines)

analysis. The remainder of the D-PBS sputum suspension was incubated with 0.2% dithiothreitol (DTT) (sigma, Poole, Dorset), placed on a bench rocker on ice for 15 minutes and filtered through pre-wet 48 μm gauze. 100 μL of this filtrate was removed for quantification of colony forming units and an additional 500 μL was removed for bacteria quantitative real time PCR (qPCR) analysis. A further 10 μL of the filtrate was removed to assess the total cell count and cell viability using a Neubauer haemocytometer. The haemocytometer was flooded with 10 μL of the filtrate mixed with 10 μL of 0.4% trypan blue (Sigma, Poole, Dorset) and all cells were counted in the four corner squares of the haemocytometer to include viable, non-viable and squamous cells. The remainder of the filtrate was then further centrifuged for 10 minutes at 790g at 4^{o}C. The DTT supernatant was removed into 300μL aliquots and stored at -80^{o}C for further mediator analysis. Following DTT supernatant removal, the cell pellet was re-suspended in a small volume of D-PBS and adjusted to make a cell suspension of 0.50 – 0.75 x10^6 cell/mL with D-PBS for cytospin preparation. 75 μL of cell suspension was placed in cytocentrifuge cups and spun at 450 rpm for 6 minutes. The slides were then air dried for 15 minutes at room temperature and stained with Rowanowski stain (0.5g Eosin, 1.5g Azure-B-thiocyanate, 10nM HEPES buffer pH7.2, DMSO). A differential cell count was obtained by counting $>$ 400 non-squamous cells on the prepared slide.

16S rDNA Based Microbiome Community

DNA Extraction From Sputum Samples: Total genomic DNA for all sputum samples was extracted using the QIAamp DNA Mini Kit assay (Qiagen, California, USA). DNA isolation from the Gram positive bacteria extraction method was followed as per the manufacturer's protocol, which involved hydrolysis of peptidoglycan cell wall layer with 20mg/ml lysozyme and incubated at 37^{o}C for 30 minutes. Further lysis was performed with Proteinase K digestion of contaminating proteins and nucleases released from the cells at 55^{o}C for 30 minutes and 95^{o}C for 15 minutes. The remainder of the extraction was done according to the "DNA Extraction from Tissue" of the manufacturer's protocol. This involved adding the cell lysate formed from the above step with 200μl absolute ethanol to the QIAamp spin column and centrifuging briefly to adsorb the DNA optimally to the column's silica gel mem-

brane. This was followed by wash steps with buffer AW1 and AW2 to remove the impurities. Finally DNA was eluted in 200μL of DNAse, RNAse free distilled water and stored at -20^{o}C.

Production of 16S rDNA bacterial amplicon library: For 454 sequencing of the bacterial community one way/unidirectional reads amplicon sequencing method was chosen, and Lib-L based method for amplicon library production was followed according to "Amplicon Library Preparation Method Manual, GS FLX Titanium Series (October 2009)".

Primer design and sequence: Primers for 454 sequencing called as fusion primers consisted of following four parts Adaptor sequence - Both forward and reverse primer have an adaptor sequence starting at the 5' prime end of the primers called as Lib-L/A and Lib-L/B respectively. These sequences allow binding of single stranded (ss) DNA to Lib-L capture beads and subsequent annealing to emulsion PCR (emPCR) and sequencing primers. Key sequence- This is a 4 base sequence "TCAG" present in both forward and reverse primer in the 5'-3' orientation downstream of the adaptor sequence. It is used as calibrator for the signal intensity produced during the sequencing flowgram cycle. Multiplex identifier (MID) sequence- For each 454 run all the samples were tagged with a unique 10 bp sequence acting as a barcode for sample identifying in downstream sequence analysis. These unique MIDs are introduced in the primer sequence immediately downstream to the key sequence. Since the 454 run performed was unidirectional sequencing only the forward primer had a MID sequence. These MIDs were chosen from the Roche's 454 set of designed MIDs for Genome Sequencer FLX titanium series (Using Multiplex Identifier (MID) Adaptors for the GS FLX Titanium Chemistry - Extended MID Set, April 2009).

Template specific primer sequence - Primers targeting the eubacterial 16S rDNA were chosen based on the following criteria:

1. The 16S oligonucleotides (oligo) were conserved amongst most eubacterial groups and at the same time non-specific to eukaryotic DNA.

2. Aim was also to include as many 16S hypervariable region for bacterial community discrimination within the optimal size PCR product for 454 sequencing.

For all 454 sequencing universal 16S primers 926F (Muyzer et al., 1995) and 1391R (Lane et al., 1985) were utilised for amplification of hypervariable regions V6 toV8.The 16S oligo sequence are given below in Table 4.2. In total 31 fusion primers consisting of 30 MID tagged 454 forward primers and a single 454 reverse primer were designed and were sourced from Sigma-Aldrich (Dorset, UK).

PCR conditions, Gel run and purification: After optimisations of PCR components yielding PCR product from most samples and the different barcoded primers, final PCR reactions were done in 50μl volume containing 1x High-Fidelity PCR buffer, 0.8μM of each forward and reverse primer, 2.5mM MgCl2, 0.25mM of each deoxyribonucleotide triphosphate (dNTP) (Promega, WI, USA), 0.8M Betaine HCl (Sigma-Aldrich, Dorset, UK), 0.08mg/ml BSA (NEB ,UK) and 0.5μl(1U) Phusion High-Fidelity DNA Polymerase (Finnzymes, Finland). 1μl DNA template was used for PCR reaction. Each PCR batch had a negative control with DNA replaced by 1μl molecular grade water. PCR cycle was performed in Dyad DNA engine involving an initial denaturation at 98^{o}C for 5 minutes and 28 cycles of 98^{o}C for 40s, 58^{o}C for 40s, 72^{o}C for 20s with a final extension at 72^{o}C for 4 minutes producing approximately 570bp long amplicon. PCR product size and purity were checked by performing gel electrophoresis as described in section (2.4.3). Most samples produced a single specific product with exception of few samples that had some non-specific products amplified as well. This might be due to the broad range of the primers utilised, producing extraneous products in presence of low amount of target gene and/or high host DNA contamination. For these samples gel purification was performed on the gel cut of the amplicon of interest using QIAquick Gel Extraction kit (Qiagen) following manufacturer's protocol. Rest of the PCR reactions were cleaned of PCR constituents and primer dimers using the Agencourt® AMPure® XP magnetic bead purification system (Beckman Coulter, USA) according to manufacturer's instructions.

DNA quantification, standardization and pooling: Subsequent quantification was via the Quant-iT^{TM} PicoGreen ® (Molecular Probes Inc., Invitrogen, USA) assay technique as per the manufacturer's instructions. Fluorescence of samples was assessed in duplicate along with 2 fold serially diluted standard DNA ranging from 100ng/μl to 1.56ng/μl, at 480 nm excitation and 520 nm emission detection, using a

FluoStar Omega Spectrophotometer (BMG Labtech, UK). Concentration (ng/μl) of dsDNA in samples was extrapolated from the standard curve and used to determine the concentration in molecules/μl. Each sample was then standardized with 1 x TE buffer to 10^9 molecules/μl. Pooled amplicon libraries were prepared using 5 μl of the standardized amplicons from each sample allocated in a 454 quarter run. Purity of the pooled amplicon library was verified using the Agilent 2100 Bioanalyzer (Agilent Technologies, UK) high-sensitivity dsDNA kit before they were dispatched for sequencing utilising the Genome Sequencer FLX Instrument titanium series (454 Life Sciences, Roche Diagnostics, UK) to Liverpool (Center for Genomic Reasearch, Liverpool, UK).

Blood Collection and Processing

A volume of 10mL of venous blood was collected by venepuncture and collected into serum gel activator (coated with silica particles to enable clotting) and EDTA plasma (coated with K2 to prevent clotting) prepared containers. These were left to stand upright for 1 hour and then centrifuged at 1700rpm for 10mins at room temperature. Venous blood was taken to measure full blood count, differential cell count.

2.4 Asthma and COPD Characteristics

Here the recoded demographic, clinical characteristic and cellular profiles, and their established clinically relevant cutoff were described. In which, the diseases severity is defined according to the lung function measurements, cellular airway inflammations based on eosinophil and neutrophil cell-counts, and bacterial colonization according to colony-forming unit (CFU) cutoff or positive culture.

2.4.1 Demographic Characteristics

Information such as age, height, weight, gender, smoking status (never smokers, ex-smokers and current-smokers) and duration of disease were recorded. Pack-year history (the amount a subject has smoked over a long time) is calculated by multiplying the number of packs of cigarettes smoked per day by the number of years

the person has smoked. In addition, body mass index (BMI) was calculated as (weight(kg)/height^2(m)).

Lung-function Measurements

Spirometry lung function measurements were carried out in accordance with the joint American Thoracic Society/European Respiratory Society (ATS/ERS) guidelines [96]. In which pre and post 400 μg salbutamol bronchodilatation forced expiratory volume in the first second (FEV_1) (which is the volume exhaled during the first second of a forced expiratory maneuver started from the level of total lung capacity), and forced vital capacity (FVC) (which is the total amount of air exhaled during the FEV test) were recorded. The best out of three consecutive blows to record the FEV_1 and the FVC was then used. The corresponding FEV_1/FVC ratio was calculated. In addition, pre and post FEV1 percentage predicted were calculated for each subject using corresponding pre or post- FEV_1, age, height and gender as reported here [96][97]. Formulated as follows;

- $Predicted\, FEV_1\% = \{\frac{FEV_1}{(0.0430*height-0.029*age-2.49)}\}*100$, for male
- $Predicted\, FEV_1\% = \{\frac{FEV_1}{(0.0395*height-0.025*age-2.60)}\}*100$, for female

Diseases Severity Based on Lung-function Measurements

The severities of the diseases were categorized using (preFEV1/FVC) ratio and postFEV$_1$PercentagePredicted (postFEV$_1$%) cut-off according to GINA and GOLD guidelines.

1. Stage I: Mild: $preFEV_1/FVC < 0.70$ and $postFEV_1\% \geq 80$
2. Stage II: Moderate: $preFEV_1/FVC < 0.70$ and $postFEV_1\% = [50 - 79]$
3. Stage III: Severe: $preFEV_1/FVC < 0.70$ and $postFEV_1\% = [30 - 49]$
4. Stage IV: Very Severe: $preFEV_1/FVC < 0.70$ and $postFEV_1\% < 30$

Bacterial Colonization

Semi-quantitative bacterial analysis was performed by colony forming units (CFU) estimation in accordance to previously described methods [97] [98]. In which 900 μL of sterile D-PBS solution was placed in 5 sterile eppendorfs labelled as 101, 102, 103, 104, 105 and serial dilutions of the 100μL DTT filtrate removed during the sputum processing procedure were made. Three 20μL drops were placed from each serial dilution onto chocolate and blood agar media. Each plate was then incubated for 24 hours in 5% CO2 at 37^{o}C. After incubation, counts were made from the dilution with <100 CFU and averaged for each of the droplets to determine a total CFU load. Positive bacterial colonization was defined as CFU greater than 107/ml sputum or positive culture [90, 98].

Visual Analogue Scale

The visual analogue scale (VAS) for the domains of i) cough ii) breathlessness (dyspnea) and iii) wheeze was used to record symptoms [99] [100]. Each subject was asked to draw on a 100mm line with 'no symptoms' at one end (0 mm) and 'the worst symptoms ever' (100 mm) at the other for each symptom domain.

Cellular Profiles

In this study, cellular profiles (such as eosinophils and neutrophils) are collected from both sputum and blood. Subjects with sputum eosinophil and neutrophil differential cell counts above 3% [64, 100] and 61% [77] were defined as eosinophilic or neutrophilic, respectively. Further stratification of the subjects into four subgroups on the basis of their sputum cell counts was also performed: pure eosinophilic (eosinophil $>$ 3% and neutrophil $\leq$ 61%), pure neutrophilic (eosinophil $\leq$ 3% and neutrophil $>$ 61%), mixed granulocytic (eosinophil $>$ 3% and neutrophil $>$ 61%), and paucigranulocytic (eosinophil $\leq$ 3% and neutrophil $\leq$ 61%). Then the proportions of asthmatics and COPD subjects in each subgroup were assessed.

2.4.2 Biological Mediators (Sputum Cytokines)

There are little direct comparisons between asthma and COPD on the cytokine profiles. In this study a number of sputum cytokines (mediators) from patients with severe asthma and COPD were measured from sputum cell-free supernatant using the Meso Scale Discovery Platform (MSDQR Gaithersburg, MD, USA) as previously described [92]. The mediators that have detectable range in over 50% of the samples in both asthma and COPD subjects were used for analysis in this study. The mediators which satisfy this criteria are: IL-1β, IL-5, IL-6, IL-6R, IL-8, IL-10, IL-13, CCL-2, CCL-3, CCL-4, CCL-5, CCL-13, CCL-17, CCL-26, CXCL-10, CXCL-11, TNFα, VEGF. These mediators can biologically be classified as TH 1 derived, TH 2 derived, and proinflammatory cytokines [48]. In which IL-5, IL-13, CCL-13, CCL-17 and CCL-26 as TH2 derived; CXCL-10 and CXCL-11 as TH1 derived; and IL-β, IL-6, IL-6R, IL-8, CCL-2, CCL-3, CCL-4, CCL-5, TNFα and VEGF as pro-inflammatory mediators. These mediators will be used to explore the biological heterogeneity of asthma and COPD in the subsequent chapters. However, in this section, how the cytokines were extract from sputum, their lower limit of detection and quantification, and the platform used to measure these cytokines (Meso Scale Discovery) will be briefed.

Laboratory Measurement of Sputum Cytokines

A wide panel of cytokines was measured using the MSD platform from the sputum of the patients according to the manufacturer's instructions. In brief, 25μL of the cytokine assay diluents was added to the plate and incubated for 30 minutes. This was followed by the addition of 25μL of sputum D-PBS supernatant and incubated for 2 hours. The plate was then washed three times with diluted wash buffer and 25μL of detection antibody was added. After a further incubation period of one hour and a repeated wash step, 150μL of read buffer was added and the plate was read.

Meso Scale Discovery Platform: an Analytes measurement

Meso Scale Discovery is analyte measurement of assays that provides a rapid and convenient method for measuring the levels of profiling cytokines within a single or small-volume sample. Although it is similar to a traditional ELISA, Meso Scale Discovery Electrochemiluminsence (MSD-ECLU) uses non-radioactive electrochemiluminescent (ECL) labels for ultra-sensitive detection. The Meso Scale Discovery MSD assay platform utilizes Ruthenium (II) trisbipyridine (4-methylsulfone) [Ru(bpy)3] that, once conjugated to the analyte, serves as the tracer in competitive assays. ECL labels generate light when stimulated by electricity in the appropriate chemical environment. High binding carbon electrodes in the bottom of microplates allow for easy attachment of biological reagents. MSD assays use ECL labels that are conjugated to detection antibodies. Electricity is applied to the plate electrodes by an MSD instrument leading to light emission by labels. MSD's assays improve sensitivity, expand the dynamic range, enable measurement of multiple analytes from a single sample (i.e. multiplexing), and work well in difficult sample types. Light intensity is then measured to quantify analytes in the sample [www.mesoscale.com].

Lower Limit of Quantification and Detection of the Cytokines

The lower limit of quantification (LLOQ) and detection (LLOD) for each kit were determined from the respective standard curves for each cytokine. The LLOQ was defined as the lowest concentration on the standard curve that satisfies the following criteria [101]:

- A measured concentration within 25% of the nominal value
- And a coefficient of variation (%CV) less than 25%

The LLOD was defined as the lowest concentration on the standard curve whose readout was greater than 2.5 standard deviations above that of the blank. In this analysis, for the samples that were below the LLQ, a value of the (LLQ/2) was assigned. The LLD and LLQ for the mediators which used in this analysis are depicted in table 2.1.

Table 2.1: Sputum mediators lower limit of detection (LLD) and quantification (LLQ)

Variable	LLQ	LLD
IL-1β (pg/ml)	2.56	1.58
IL-5 (pg/ml)	0.64	0.476
IL-6 (pg/ml)	0.64	0.516
IL-6R (pg/ml)	0.64	0.443
IL-8 (pg/ml)	2.56	0.617
IL-10 (pg/ml)	3.2	2.17
IL-13 (pg/ml)	16	6.75
CXCL-10 (pg/ml)	12.8	3.2
CXCL-11 (pg/ml)	3.2	1.24
CCL-2 (pg/ml)	3.2	2.5
CCL-3 (pg/ml)	16	13.2
CCL-4 (pg/ml)	16	6.05
CCL-5 (pg/ml)	3.2	1.21
CCL-13 (pg/ml)	16	13.4
CCL-17 (pg/ml)	0.64	8.93
CCL-26 (pg/ml)	3.2	0.932
TNFα (pg/ml)	0.64	0.31
VEGF (pg/ml)	400	125

2.4.3 Descriptive Statistical Analysis

Subjects from severe asthma (according to GINA) and COPD (according to GOLD) were combined based on the common measurements that exist in both studies at stable visit. However, prior to merging, consistencies of all common measurements across the two studies were thoroughly checked using appropriate statistical and graphical techniques. Distributions of continuous variables were assessed, and natural logarithm transformations of positive skewed variables were used in subsequent analyses, as appropriate. As distributions of all sputum cytokines were positively skewed, analysis was carried out on their logarithmic values throughout. Thus, descriptive data analysis was performed separate on the demographic, clinical and biological characteristics across asthma and COPD.

Discovering and understanding the entire relationship in dimensional data is quite problematic, particularly when the underlying structures are largely unknown. Therefore, to understand the overall patterns of the data across asthma and COPD, principal component analysis (PCA) was performed separately on demographic and clinical characteristics, and then on sputum mediators, and are presented graphically across their first two principal components. PCA is a linear combination of observed variables to form new independent latent variables called components. In addition, correlation matrices of the cytokines are reported graphically as a heatmap. Furthermore, the variables (cytokines) subgroups (based on their correlation matrices) were investigated whether the cytokines were arranged into homogeneous and biologically meaningful subgroups using the Clustofvar [101] R package procedure. This procedure organises the set of variables into hierarchical clusters, and the results are presented graphically as dendrogram showing cytokines in each subgroup, and the distance between the subgroups.

2.5 Descriptive Results

2.5.1 Demographic and Clinical Characteristics

Preliminary descriptive analysis (group comparisons using analysis of variance, and chi-square test for continuous and categorical variables, respectively) on the demographic and clinical characteristics was performed across asthma and COPD subjects at disease level. Normal data were presented as arithmetic mean with standard error of the mean (SEM), log-transformed data as geometric mean with corresponding 95% confidence interval (CI). The χ^2 test or Fisher exact test was used to compare proportions, and 1-way ANOVA was used to compare means across multiple groups. Non-normal data as median with 1^{st} and 3^{rd} interquartile range, and Kruskal-Wallis test was used to compare subgroups. The results are depicted in table 2.2 on page 44.

Table 2.2: Statistical summaries of demographic and clinical characteristics across asthma and COPD at stable state that shows the similarities and differences between the two diseases

Variable	Asthma	COPD	P-value
Male (%)	43 (50)	53 (70.7)	0.008
Current or Ex- smokers [n (%)]	32 (37.2)	72 (96.0)	< 0.0001
Pack -year history^{s}	4.6 (2.98 - 7.26)	40 (34.46 - 46.39)	<0.0001
Age (years)$^{+}$	54 (1.3)	69 (1.1)	<0.0001
Duration of Disease (years)	21 (16.4 - 26.5)	5 (4.12 - 6.55)	<0.0001
BMI (kg/m2)$^{+}$	30.4 (0.8)	25.7 (0.5)	<0.0001
Exacerbation number of steroids$^{\delta}$	3 (0.23)	4 (0.31)	0.007
Maintenance prednisolone dose use [n (%)]	52 (60.5)	8 (10.7)	<0.0001
Daily Prednisolone dose (mg)*	10 (7.5 - 15)	5 (5 - 5)	0.002
Daily Inhaled Corticosteroid dose (mcg/day)^{*a}	1600 (100 - 2000)	1200 (800 - 2000)	0.05
Pre FEV_1 (L)$^{+}$	2.15 (0.1)	1.28 (0.1)	<0.0001
Pre FEV1/FVC ratio (%)$^{+}$	67.6 (1.5)	49.8 (1.6)	<0.0001
Pre FEV1 Predicted (%)$^{+}$	74.6 (2.4)	45.4 (2.1)	<0.0001
Post FEV_1 (L)$^{+}$	2.32 (0.09)	1.32 (0.06)	<0.0001
Post FEV_1 Predicted (%)$^{+}$	79.8 (2.4)	47.1 (2.1)	<0.0001
Sputum Neutrophil count (%)$^{+}$	63.2 (2.5)	69.7 (2.5)	0.07
Sputum Eosinophil count (%)	2.1 (1.38 - 3.1)	1.4(0.98 - 1.93)	0.14
Sputum Macrophage count (%)	16.7 (13.41 - 20.78)	16.2 (13.3 - 19.8)	0.84
TCC ($x10^6 cells/gsputum$)	1.64 (1.28 - 2.11)	3.34 (2.53 - 4.41)	<0.0001
Blood Eosinophil x109/L	0.23 (0.19 - 0.29)	0.22 (0.19 - 0.26)	0.63
Blood Neutrophil x109/L^{+}	5.81 (0.2)	5.59 (0.2)	0.5
CFU >107/ml or positive culture (n[%])	16 (18.6)	30 (40)	0.003
VAS-cough (mm)$^{+}$	34 (2.7)	44 (3.4)	0.021
VAS-dyspnoea (mm) $^{+}$	34 (2.8)	46 (3.0)	0.004

Definition of abbreviations: VAS= Visual Analogue Score; BMI= Body Mass Index; FEV1=Forced Expiratory Volume in the First Second; FVC=Forced Vital Capacity; TCC=Total sputum cell count CFU= colony forming units. Data presented as geometric mean (95% CI) unless stated;$^{+}$Mean (standard error of mean (SEM)); *median (1st and 3rd quartile); Dose for only those subjects prescribed daily prednisolone; ^{s}Pack-year history of current and ex-smokers; ^{a}beclomethasonedipropionate equivalent; $^{\delta}$= Total number of times a patient exacerbated and took high dose of steroids for at least three days in the last 12 months.

As we observe in the table above, there are significant differences between asthma and COPD in many characteristics. For example, COPD subjects are more likely to be men and older, lower in BMI, high in symptom visual analogue scores, have low lung function measurements, and with high proportion of bacterial colonisation compared to the asthmatic subjects. However, there are no significant differences between the two diseases in sputum and blood eosinophil and neutrophil cell-counts.

In addition, the subjects in each disease were categorized according to their lung severities (airflow obstruction) based on lung function measurements. Therefore, 23.3% of asthmatic and 2.8% of COPD subjects have mild, 24.4% of asthmatic and 31.9% of COPD have moderate, 7.0% of asthmatic and 37.5% of COPD have severe, and 2.3% of asthmatic and 18% of COPD have very severe airflow obstruction.

Furthermore, the subjects in both diseases were separately assessed their inflammation based on their cell-counts. In which 25.9% of asthmatic and 13.7% of COPD subjects are pure eosinophils; and 37.6% of asthmatic and 52.0% of COPD are pure neutrophils; 15.3% of asthmatic and 17.8% of COPD are pausi, and 21.2% asthmatic and 16.4% of COPD are mixed granulocytic.

Moreover, principal component analysis was performed on the above continuous demographic and clinical characteristics to understand the overall patterns of the asthma and COPD overlap, and graphically presented across their first two principal component scores in figure 2.2 on page 45.

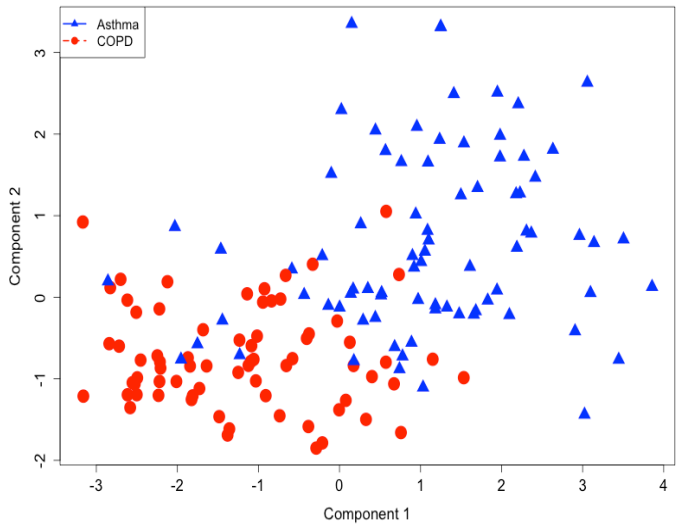


Figure 2.2: Clinical characteristics across asthma and COPD at stable state which displayed on the first two principal components scores

As shown in the figure above, asthma and COPD are quite distinctive on their demographic and clinical characteristics. This means that the two diseases can be easily distinguished using the combination of demographic and clinical characteristics.

2.5.2 Sputum Biological Mediators

To understand patterns of the sputum cytokines across asthma and COPD, descriptive analysis was preformed and results are presented as summary statistics in table 2.3 on page 46.

Table 2.3: Statistical summaries of sputum mediators across asthma and COPD at stable state that represent the similarities and differences between the two diseases

Variable	Asthma	COPD	P-value
IL-1β (pg/ml)	70.3 (53 - 93.1)	73.5 (47.1 - 114.6)	0.86
IL-5 (pg/ml)	2.7 (1.9 - 4.0)	1.1 (0.8 - 1.5)	0.001
IL-6 (pg/ml)	42.8 (30.5 - 60.1)	439.2 (320.5 - 601.8)	<0.0001
IL-6R (pg/ml)	243.4 (195.1 - 303.7)	147.8 (119.3 - 183.1)	0.002
IL-8 (pg/ml)	3118 (2314 - 4201)	4390 (3372 - 5715)	0.098
IL-10 (pg/ml)	0.73 (0.5 - 1.1)	1.0 (0.6 - 1.7)	0.27
IL-13 (pg/ml)	8.1 (6.4 - 10.3)	3.4 (2.5 - 4.6)	<0.0001
CCL-2 (pg/ml)	284.3 (226.6 - 356.7)	573.1 (455.3 - 721.4)	<0.0001
CCL-3 (pg/ml)	30.6 (22.9 - 40.7)	67.1 (53.2 - 84.7)	<0.0001
CCL-4 (pg/ml)	359.5 (247.0 - 523.3)	958.3 (778.5 - 1179.7)	<0.0001
CCL-5 (pg/ml)	8.7 (6.9 - 11.0)	3.3 (2.6 - 4.1)	<0.0001
CCL-13 (pg/ml)	19.2 (14.7 - 25.0)	28.1 (21.3 - 37.0)	0.052
CCL-17 (pg/ml)	25.9 (19.5 - 34.6)	20.3 (15.1 - 27.2)	0.24
CCL-26 (pg/ml)	9.9 (6.8 - 14.3)	2.9 (2.0 - 4.1)	<0.0001
CXCL-10 (pg/ml)	726.9 (526 - 1004.7)	277.9 (205.3 - 376.3)	<0.0001
CXCL-11 (pg/ml)	57.2 (39.7 - 82.5)	11.6 (7.6 - 17.9)	<0.0001
TNFα (pg/ml)	3.2 (2.3 - 4.5)	5.4 (3.3 - 8.9)	0.093
VEGF (pg/ml)	1427 (1214 - 1678)	1284 (1129 - 1461)	0.33

Data presented as geometric mean with corresponding 95% confidence interval (CI).

As depicted in the table above, IL-5, IL-6R, IL-13, CCL-2, CCL-5, CCL-26, CXCL-10, and CXCL-11 are significantly higher in asthmatic compare to COPD subjects. In contrast, IL-6, CCL-3 and CCL-4 are significantly higher in COPD subjects compare to asthmatic. However, there are no significant differences between the two diseases in IL-1β, IL-8, IL-10, CCL-13, CCL-17, TNFα and VEGF. To understand the overall patterns of the mediators across asthma and COPD, their z-scores (standardized value) are plotted in figure 2.3 on page 47.

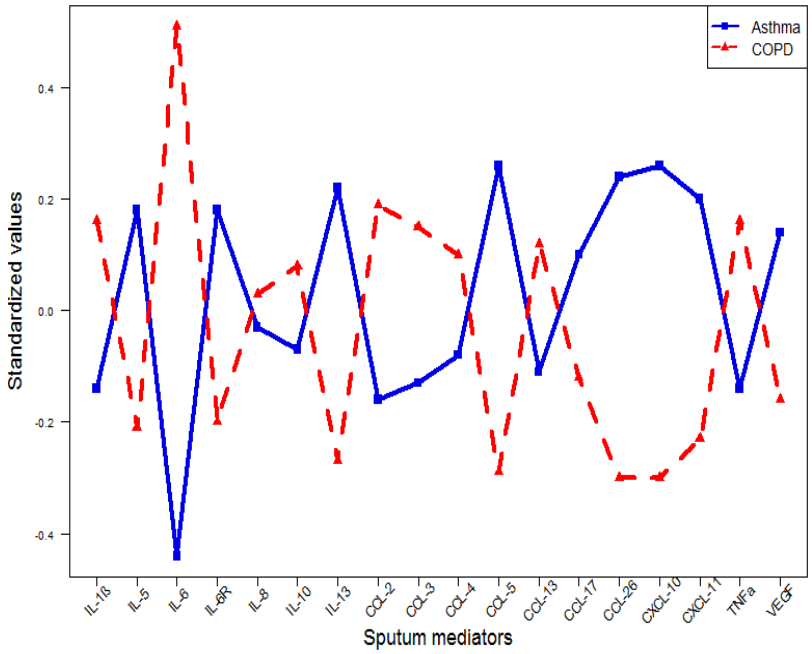


Figure 2.3: Pattern of sputum mediators across asthma and COPD at stable state

In addition, to understand the overlap and distinction between asthma and COPD at biological level, principal component analysis was performed to the sputum mediators, and graphically displayed in figure 2.4 on page 48 across the first two PCA scores.

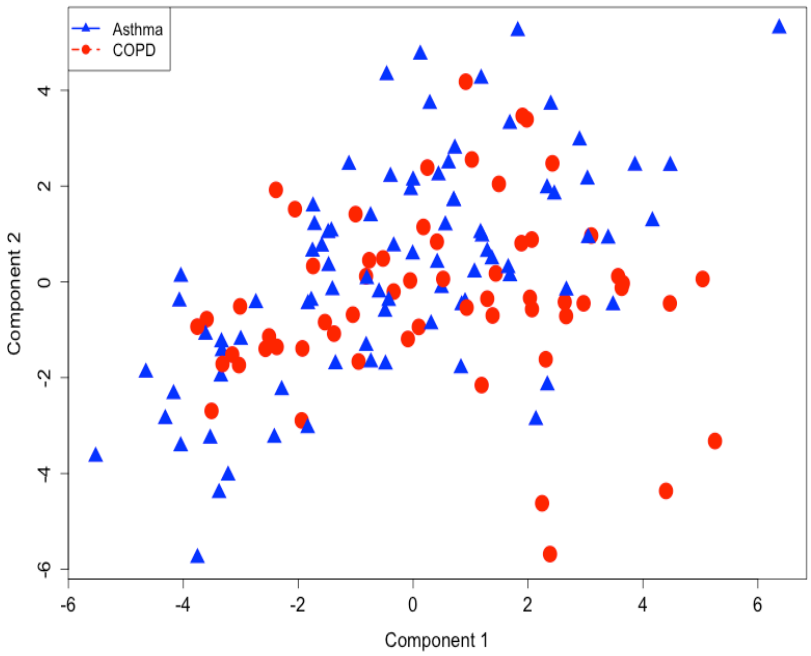
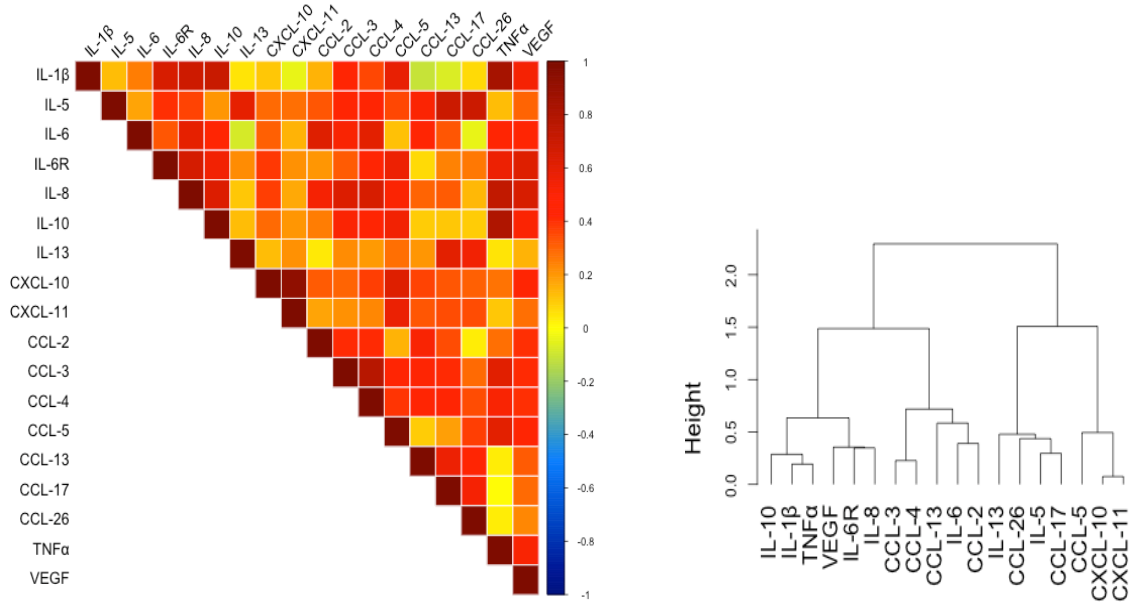


Figure 2.4: Sputum mediators across asthma and COPD at stable state presented using the first two principal component scores

In figure 2.4 above, there is an evident overlap between asthma and COPD subjects along the first two principal components direction of the cytokines, which is in contrast to the patterns observed across the diseases with respect to the demographic and clinical characteristics in figure 2.2 on page 45. From this descriptive analysis, it is difficult to quantify the proportion of overlap between the two diseases with respect to the biological mediators. Therefore, it requires further investigation using appropriate statistical method in order to identify the common and distinctive biological subgroups of asthma and COPD.

Furthermore, correlations among the mediators were displayed as heatmap in figure 2.5(a); and further variables' subgroups, using Clustofvar version 0.8 R package [101], were assessed to investigate for hidden structures/patterns among the mediators. Thus, based on the patterns of their correlations, the mediators were partitioned into distinctive subgroups, and depicted in figure 2.5(b). Correlation is the normalization of the covariance by the square-root of their variance product that range from -1 (strong negative correlation) to +1 (strong positive correlation).

Covariance between two mediators shows the extent that the two mediators spread together.



(a) (b)

Figure 2.5: Sputum mediators at stable state: (a) correlations matrix and (b) subgroups. Heatmap colours: Dark red indicates strong positive correlation; dark blue for strong negative correlation; light red for weak positive correlation; light blue for weak negative correlation; and yellow represents no correlation.

As we observe in the figures above, there are strong correlations between the mediators; and about three to four subgroups also exist among the mediators. For example, IL-1β, IL-8, IL-10, TNFα and VEGF are aggregated together; IL-6, CCL-2, CCL-3, CCL-4 and CCL-13 created a group; most of the T_H2 mediators (IL-5, IL-13, CCL-17 and CCL-26) formed another subgroup, and these T_H1 mediators (CXCL-10, CXCL-11 and CCL-5) were also grouped together. These patterns show that there are hidden structures among the mediators, which should be accounted in further analysis for better understanding of the two diseases.

2.6 Summary

In this chapter, an explanatory data analysis was performed on the available characteristics across asthma and COPD at stable state. It was clearly shown that asthma and COPD are distinctive with respect to clinical and demographic characteristics except in their cell-counts (eosinophils and neutrophils). However, there is a considerable overlap between the two diseases with respect to their sputum cytokines,

which seeks further investigation using appropriate statistical techniques to identify the common and distinctive biological subgroups. In addition, strong correlations were observed among the cytokines, and appeared to create cytokines' (variables') subgroups. Thus, these patterns need to be accounted whilst modeling (clustering) the biological heterogeneity of the diseases.

Chapter 3

Modeling Population Heterogeneity: a Simulation Study

3.1 Objects

The objective of this chapter is to identify an appropriate statistical technique for modeling the biological heterogeneity of asthma and COPD using simulation study. An artificial data were simulated which have similar internal patterns as the asthma and COPD cytokines, but with known class membership. The performance of several approaches were assessed using the simulated data. The method which performed best in the simulation study was applied to the real asthma and COPD cytokines studies in the subsequent chapter, in order to identify the common and distinctive biological clusters of the diseases.

3.2 Introduction

In the previous explanatory analysis (chapter 2) it has been shown that asthma and COPD are clearly overlapped with respect to the biological mediators (sputum cytokines), which requires further investigation. In addition, the mediators appeared to be strongly correlated and created subgroups based on their correlation matrix. Thus, the internal patterns (correlations) of these mediators need to be accounted (integrated) whilst clustering the subjects, in order to identify the optimal biological clusters of both diseases.

It's not fully understood yet that whether a standard technique (i.e. using the actual correlated sputum mediators as input into a clustering) may lead to the identification of the optimal clusters (although this approach ignores the correlation between variables within a cluster). On the other hand, whether a dimensionality

reduction using factor analysis (i.e. representing the actual cytokines by low dimensional latent variables that capture the internal structures of the mediators, and use as input into a clustering algorithm) may improve the clusters partitions.

Therefore, this chapter focused on a simulation study in which an artificial data were simulated. The variables in the simulated data have similar patterns as the sputum cytokines, but the observations have known class membership. The artificial variables were represented using several variables such as latent or highest-loading variables (derived from factor analysis), and were independently used as input variables into the k-means clustering algorithm. Factor analysis has been chosen versus Principal component analysis for data reduction as the main aim of this study is on a theoretical solution of the underlying structure of the cytokines which is uncontaminated by unique and error variability [102]. The performance of each method was assessed in identifying the known simulated clusters. In addition, the consequence (technical issues/bias) of using inappropriate input variables into the clustering algorithm was discussed. The approach which led to the identification of the optimal clusters is applied to asthma and COPD sputum cytokines in subsequent chapters.

Proposed Input Variables to Clustering

1. All observed variables (standard approach)
2. Factor scores (latent variables)
3. Highest-loading observed variables
4. Observed variables with highest error-terms

3.2.1 Designing a Simulation Study

Artificial data was simulated to compare the performance of the above approaches in using as input variables into a clustering algorithm in order to identify the optimal simulated clusters. In total 1000 observations were simulated from four known clusters using mixture model. The observations have measurements on 20 variables in which some of the variables are strongly correlated within a cluster.

Modeling and Analysis Strategies of the Simulation Study

1. First descriptive analysis such as correlations among the variables, using heat-map, and variables' subgroups were assessed.

2. Factor analysis was performed, and unrotated and rotated (using varimax) factor loadings were calculated.

3. Factor scores (latent variables) corresponding to each factor were estimated for each subject (observation).

4. Highest-loading variables which represent each factor were identified.

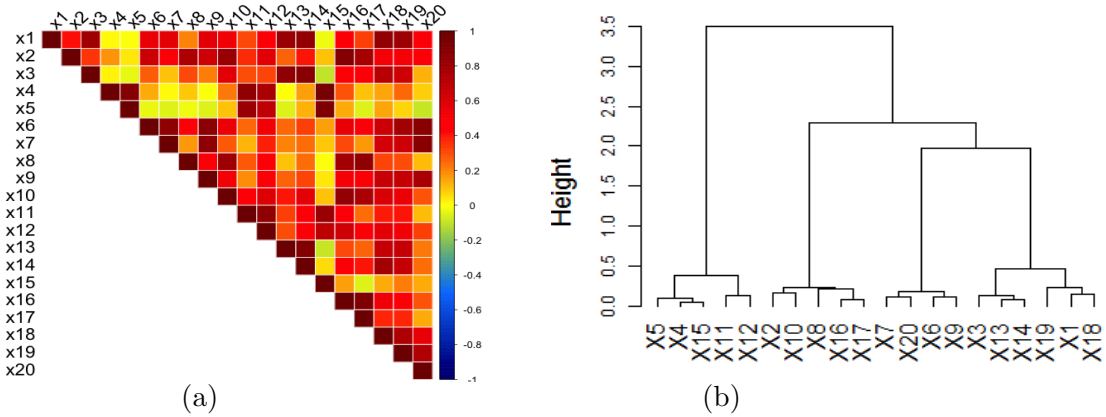
5. Variables which have relatively high error-terms were identified.

6. K-means cluster analysis was performed separate to "All simulate variables"; "Factor scores"; "Highest-loadings variables" and "Variables with highest error-terms".

7. The performance of factor analysis for uncorrelated data was also assessed

3.2.2 Descriptive Analysis of the Simulated Data

The correlation matrix of the simulated variables is displayed as heatmap in figure 3.1(a). In addition, the variables subgroups based on their internal correlations (patterns) are depicted in figure 3.1(b) on page 54. Variables which are strongly correlated appeared to aggregate (grouped) together, and the entire variables created four subgroups based on their internal structures (correlations); in which X_4, X_5, X_{11}, X_{12} and X_{15} are aggregated together; X_2, X_8, X_{10}, X_{16} and X_{17} are grouped together; X_6, X_7, X_9 and X_{20} are clustered together; and X_1, X_3, X_{13}, X_{14}, X_{18} and X_{19} formed another subgroup.



(a) (b)

Figure 3.1: Simulated variables: (a) correlation matrix and (b) subgroups. Heatmap colours: dark-red represented for strong positive correlation; dark-blue for strong negative correlation; yellow for no correlation; light-red and light-blue for weak positive and negative correlation, respectively.

In addition, the profile of the variables across the simulated clusters (subgroups) is depicted in figure 3.2 on page 54; in which X_1, X_3, X_{13}, X_{14}, X_{18} and X_{19} are elevated in cluster 1; X_6, X_7, X_9 and X_{20} in cluster 2; X_4, X_5, X_{11}, X_{12} and X_{15} in cluster 3; and X_2, X_8, X_{10}, X_{16} and X_{17} in cluster 4 compared to the other clusters.

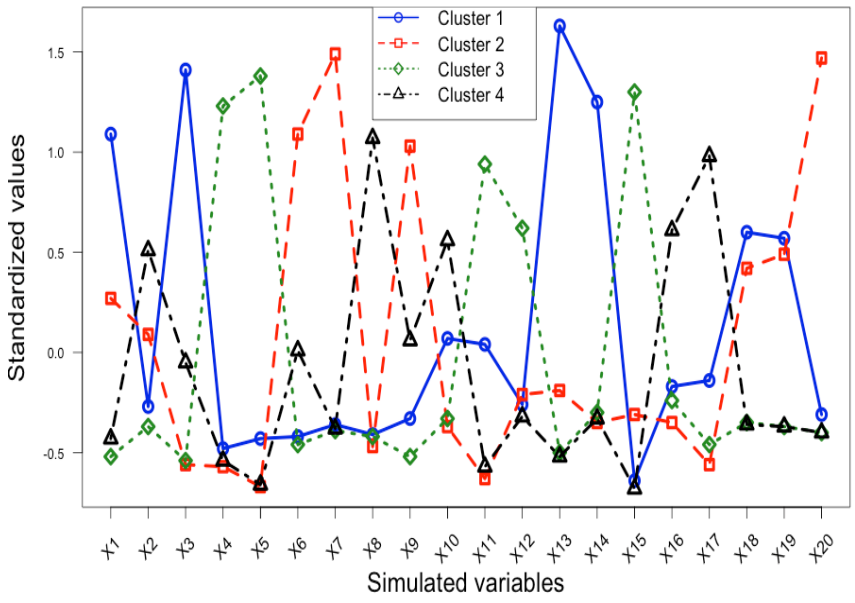


Figure 3.2: Patterns of simulated variables across the clusters

3.3 Factor Analysis

The twenty simulated variables were reduced into four independent factors using factor analysis. Factor analysis is a statistical technique which uses for investigating whether p dimensional observed variables (Y_1, Y_2, ...,Y_p) are linearly related to (function of) a smaller number of k unobservable (latent) factors (F_1, F_2, ..., F_k).

Mathematically it's formulated as follows:

$$\begin{aligned} Y_1 &= \beta_{11}F_1 + \beta_{12}F_2 + \cdots + \beta_{1k}F_k + \epsilon_1 \\ Y_2 &= \beta_{21}F_1 + \beta_{22}F_2 + \cdots + \beta_{2k}F_k + \epsilon_2 \\ &\vdots \\ Y_p &= \beta_{p1}F_1 + \beta_{p2}F_2 + \cdots + \beta_{pk}F_k + \epsilon_p \end{aligned} \tag{3.1}$$

Where: Y_p are the observed variables; F_k are latent (unobserved) factors; and $p > k$. The parameters β_{ij} are the factor loadings which represents the relationship between the observed variables (Y_p) and unobserved factors (F_k). For example, β_{11} is called the loading of variable Y_1 on factor F_1, and so forth.

Assumptions of Factor Analysis

The unobservable factors (F_k) are independent of one another, and $E(F_k) = 0$ and $Var(F_k) = 1$; the error term (ϵ_p) are also independent of one another, and E(ϵ_p) = 0 and Var(ϵ_p) = σ_p^2. F_k and ϵ_p are independent in which Cor(F_k, ϵ_p)=0.

Factor analysis can be represented graphically as path diagram. For example, assuming that five observed variables (Y_1, Y_2, Y_3, Y_4 and Y_5) are represented by two independent factors, F_1 and F_2, and are depicted in figure 3.3 on page 56. The betas (β_{ij}) represented the relationship (correlation) between the observed variables (Y_i) and latent factors (F_j), and e_i are the error terms of the observed variables (Y_i) which not explained by the factors (F_j).

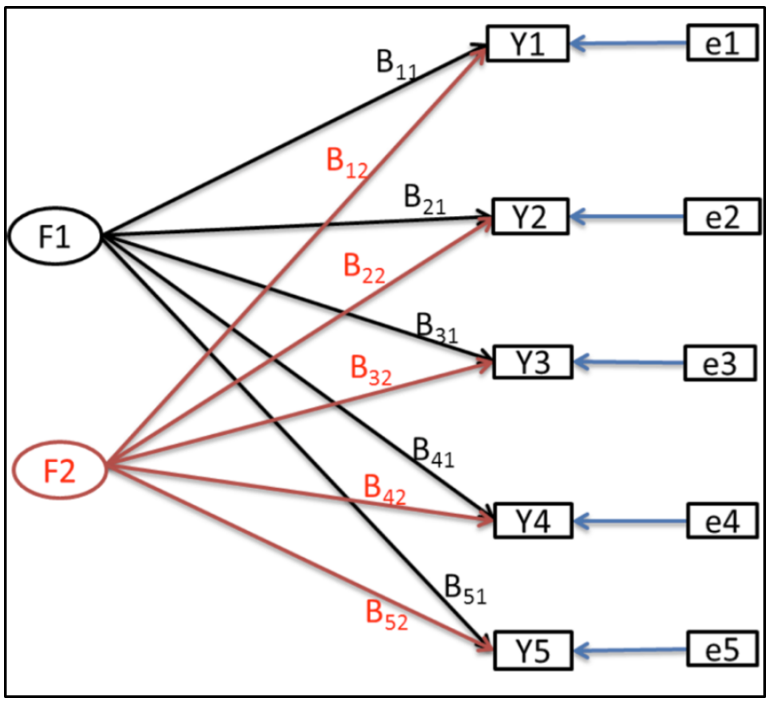


Figure 3.3: Path-diagram of factor analysis. Y_1 to Y_5 are observed variables; F_1 and F_2 are factors; B_{11} to B_{51} represent factor loadings; e_1 to e_5 are error terms.

Communalities and Error-terms

Communality (shared variance) is the sum of squared loadings for a variable across factors. Whereas, the error-terms (specific variance) is the part of the variance of the observed variable (Y_i) that is not explained by the common factors ($1 - communality$), which is known as uniqueness.

The communalities and error terms of the observed variables (Y_i) can be estimated from the latent factors as follows:

Since $Y_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + \cdots + \beta_{ik}F_k + (1)e_i$. This equation consists of two parts: $\{(\beta^2)_{i1} + (\beta^2)_{i2} + \cdots + (\beta^2)_{ik}\}$ is the communality (shared variance); and σ_i^2 is the specific (unique) variance (i.e. the error-terms).

3.3.1 Factor Loadings

Factor loadings are matrix of regression-like weights used to estimate the unique contribution of each factor to the observed variables [102]. The sum of squared loadings ($\Sigma_i^p \beta_{ik}^2$) can be interpreted as the contribution of factor F_k in explaining the sum of the observed variables (Y_p) shared variances.

3.3.2 Factor Scores or Latent Variables

A factor score (latent variable) is a numerical value that indicates an observation's relative spacing or standing on a latent factor [102]. Factor scores (FS) are independent to each other with mean of zero and standard deviation of one. The scores can be calculated for each observation as follows:

$$FS = Z(S^{-1}\beta) \tag{3.2}$$

Where: Z is the standardised value (z-score) of the observed variables; S^{-1} is the inverse correlation matrix of the observed variables; and β is factor loading matrix.

Example

Unrotated Factor Loadings of the Simulated Data

Factor analysis (principal factor) was applied to the twenty simulated variables, and subsequently the variables were reduced to four independent factors.

Since determination of possible number of factors to extract using goodness of fit was difficult to meet the assumptions required to the significance tests, therefore heuristics methods were used. Factors were retained based on two criterion:- such that eigenvalues (amount of original shared variance accounted) greater than one (i.e. a factor variance should at least represent more than a single variable shared variance), and based on screeplot (i.e. factors above the break in the curve) [102].

The retained factors of the simulated data are depicted in table 3.1 on page 59. The squared of the factor loading is the percentage of shared variance of the observed variable, which is explained (accounted) by the factor. For example, β_{11}

is the loading of X_1 on factor 1 (i.e. $\beta_{11} = 0.78$), and its squared value is 0.61 { i.e. $(\beta_{11})^2 = (0.78)^2 = 61\%$}, which means that 61% of the shared variance of X_1 is explained (accounted) by factor 1. The total variance of a single observed variable which are explained by all the retained factors is calculated as the sum of the squared factor loadings (row wise). For example, the total variance of X_1 explained by these retained four factors is calculated as $(0.78)^2 + (-0.28)^2 + (0.44)^2 + (0.17)^2 = 0.91$, which is called the communality (C^1). Whereas, the variance of X_1 not explained by these factors is one less than the communality (1 - 0.91 = 0.09), known as uniqueness (error-term). The total shared variance explained by each factor is equivalent to the eigenvalue which is calculated as the sum of the squared factor loadings (column wise). For example, the total variance of all variables explained by factor 1 is calculated as $\{(0.78)^2 + (0.78)^2 + (0.66)^2 + (0.35)^2 + (0.22)^2 + (0.79)^2 + (0.62)^2 + (0.61)^2 + (0.75)^2 + (0.81)^2 + (0.58)^2 + (0.77)^2 + (0.61)^2 + (0.74)^2 + (0.28)^2 + (0.81)^2 + (0.66)^2 + (0.85)^2 + (0.85)^2 + (0.61)^2 = 9.27\}$. The total shared variances of the variables explained by the retained four factors is calculated as the summation of column c^1 (communalities) { i.e. $(0.91+0.85+0.92+0.94+0.93+0.94+0.92+0.89+ 0.87+0.85+0.93+0.91+0.92+0.92+0.96+0.92+0.90+0.83+0.86+0.91) = 18.8\}$. Therefore, the percentage of the shared variances explained by the first factor is 9.27/18.8= 0.513. This means that about 51.3% of the total shared variance of the observed variables is accounted by the first factor.

Table 3.1: Unrotated factor loadings of the simulated data

Variable	Factor1	Factor2	Factor3	Factor4	C^1	U^2
X_1	0.78	-0.28	0.44	0.17	0.91	0.09
X_2	0.78	-0.07	-0.38	-0.28	0.85	0.15
X_3	0.66	-0.23	0.60	-0.27	0.92	0.08
X_4	0.35	0.90	0.01	0.08	0.94	0.06
X_5	0.22	0.93	0.10	0.09	0.93	0.07
X_6	0.79	-0.19	-0.38	0.36	0.94	0.06
X_7	0.62	-0.25	-0.30	0.61	0.92	0.08
X_8	0.61	-0.06	-0.44	-0.56	0.89	0.11
X_9	0.75	-0.27	-0.37	0.32	0.87	0.13
X_{10}	0.81	-0.03	-0.13	-0.41	0.85	0.15
X_{11}	0.58	0.75	0.16	-0.02	0.93	0.07
X_{12}	0.77	0.56	-0.06	0.01	0.91	0.09
X_{13}	0.61	-0.24	0.70	-0.03	0.92	0.08
X_{14}	0.74	-0.10	0.59	-0.11	0.92	0.08
X_{15}	0.28	0.91	-0.04	0.24	0.96	0.04
X_{16}	0.81	0.03	-0.27	-0.45	0.92	0.08
X_{17}	0.66	-0.12	-0.29	-0.60	0.90	0.10
X_{18}	0.85	-0.14	0.25	0.17	0.83	0.17
X_{19}	0.85	-0.13	0.16	0.30	0.86	0.14
X_{20}	0.61	-0.24	-0.27	0.64	0.91	0.09
Eigenvalue	9.27	3.86	2.48	2.46		

C^1 = Proportion of total variation accounted by the common factors (common variance)
U^2 = Proportion of total variation not accounted by the common factors (unique variance)

3.3.3 Varimax Rotation in Factor Analysis

After factor analysis was performed, the retained factors (e.g. table 3.1) can be rotated using several approaches, the most common one is using varimax rotation. Varimax is an orthogonal rotation of factors that maximise the variance of factor loadings by making high-loadings higher and low-loadings lower in each factor [102, 103]. In other word, it encourages the detection of factors related to few variables and discourages the detection of factors influencing all variables [104]. This procedure improves interpretation of the factors as the first unrotated factors (e.g. table 3.1 on page 59) usually does not reveal a clear pattern of the loadings; in which some variables' loadings are very similar across the factors, and quite difficult to figure out

which factor represents which observed variables. Despite the fact that the varimax rotation improves interpretation, it does not use for improvement of model fitting as all orthogonally rotated solutions are mathematically equivalent to one another [102].

Varimax Rotation is formulated as follows:

Varimax rotation can be accomplished by maximising the variance of the loadings on factors. The variance of the K^{th} factor can be computed as follows:

$$S_k^2 = \frac{K \sum_{j=1}^{p} \left(\beta_{jk}/\omega_j^2\right)^2 - \left(\sum_{j=1}^{p}(\beta_{jk}^2/\omega_j^2)\right)^2}{K^2} \tag{3.3}$$

Where, $\omega_j^2 = \sum_{i=1}^{k} \beta_j$ is the communality of the j^{th} variable; K is the number of retained factors; p is the number of observed variables; β_{jk} is the loading of variable j on factor k. Using this expression of the variance of the loading on the k^{th} factor, it can be maximising the following:

$$V = \sum_{k=1}^{K} S_k^2 \tag{3.4}$$

This is an iterative process where two factors rotate at a time, holding other factors constant, until the increase in the overall variance V drops below the present value.

The original unrotated factors (β_{un}) can be rotated (β_r) orthogonally as follows:

$$\beta_r' = \beta_{un}\,\Gamma$$

where

$$\Gamma = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

Where θ is any angle in degree which maximizes the variance V.

Example

Rotated Factor Loadings of the Simulated Data

The retained factors of the simulated data (table 3.1) are rotated using varimax rotation and depicted in table 3.2 on page 62. As we observe in the simulated data, the variables appeared to create patterns across the factors, in which a variable loaded in a single factor rather than in multiple factors, which is in contrast to the pattern shown in the unrotated factors (table 3.1 on page 59). For example, X_3 in the unrotated factors loaded similarly in factor 1 and factor 3; however, when the factors rotated using varimax procedure, it is clearly loaded only in factor 1, table 3.2 on page 62. This means that most of the shared variance of X_3 is explained by factor 1 than the remaining factors. In addition, varimax rotation has an advantage to identify a representative observed variable (highest loading variable) for each factor. For instance, X_{13} has the highest loading in factor 1; X_8 in factor 2; X_7 in factor 3 and X_{15} in factor 4. This means that all the four factors could be represented by these four observed variables. However, it is not always guaranteed that the entire information exists in the factors are captured using only the highest loading observed variables (X_7, X_8, X_{13}, and X_{15}).

Table 3.2: Varimax rotated factor loadings of the simulated data

Variable	Factor 1	Factor 2	Factor 3	Factor 4	C^1	U^2
X_1	0.83	0.13	0.45	0.02	0.91	0.09
X_2	0.16	0.81	0.39	0.12	0.85	**0.15**
X_3	0.91	0.29	-0.01	-0.02	0.92	0.08
X_4	0.00	0.09	0.01	0.96	0.94	0.06
X_5	-0.01	-0.04	-0.09	0.96	0.93	0.07
X_6	0.16	0.37	0.88	0.09	0.94	0.06
X_7	0.12	0.07	**0.95**	0.02	0.92	0.08
X_8	0.03	**0.93**	0.12	0.03	0.89	0.11
X_9	0.15	0.38	0.84	0.00	0.87	0.13
X_{10}	0.38	0.80	0.21	0.15	0.85	**0.15**
X_{11}	0.29	0.22	0.05	0.89	0.93	0.07
X_{12}	0.25	0.42	0.30	0.76	0.91	0.09
X_{13}	**0.95**	0.06	0.10	0.00	0.92	0.08
X_{14}	0.91	0.23	0.11	0.15	0.92	0.08
X_{15}	-0.09	-0.05	0.11	**0.97**	0.96	0.04
X_{16}	0.26	0.88	0.21	0.20	0.92	0.08
X_{17}	0.20	0.92	0.07	-0.01	0.90	0.10
X_{18}	0.69	0.25	0.52	0.17	0.83	**0.17**
X_{19}	0.61	0.20	0.64	0.18	0.86	**0.14**
X_{20}	0.13	0.03	0.94	0.03	0.91	0.09
Eigenvalue	4.63	4.56	4.53	4.33		

C^1 = Proportion of total variation accounted for by the common factors (common variance)
U^2 = Proportion of total variation not accounted by the common factors (unique variance). These bold loadings are for these highest loading variables in each factor; and these bold error-terms (U^2) are for these variables in which their shared variance was not relatively explained well by these retained four factors.

3.4 Cluster Analysis

Cluster analysis is an unsupervised statistical technique that uses to uncover unknown structures among heterogeneous population (figure 3.4(A)). These heterogeneous population can be partitioned into distinctive subgroups based on the similarities and differences of the characteristics of the population (figure 3.4(B)).

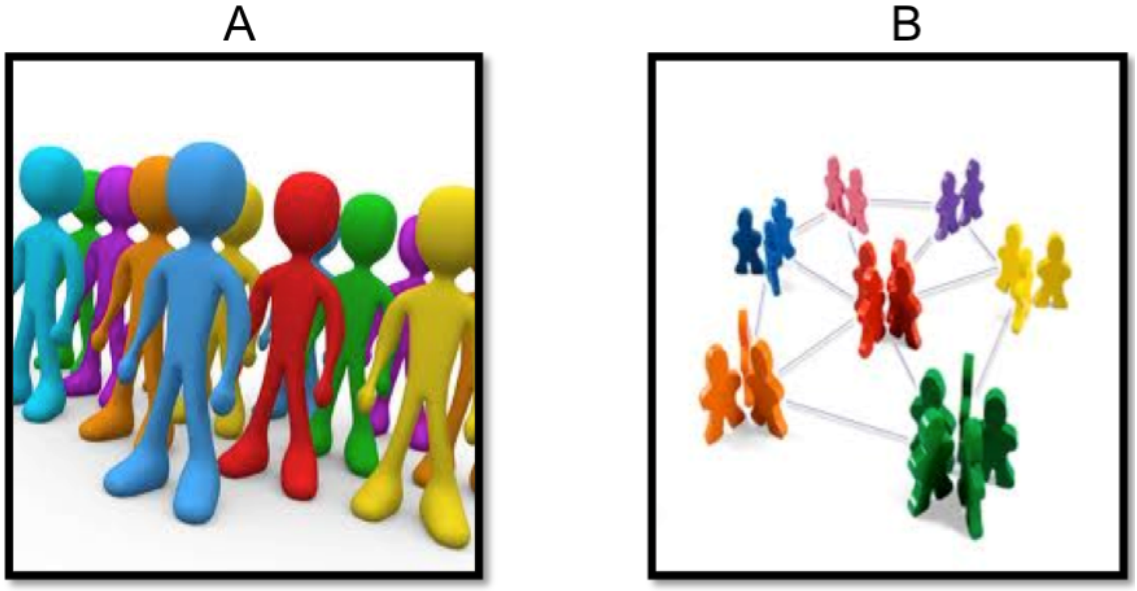


Figure 3.4: (A) Heterogeneous population; (B) Homogeneous subgroups

Cluster analysis is a useful technique to glean novel subgroups (clusters), and to generate a new research hypothesis for supervised techniques such as linear regression, discriminant and pathway analyses. There are two common types of clustering techniques; i.e. heuristic and model-based approaches. The heuristic approach is a distance based, which comprises several clustering algorithms such as hierarchical and k-means. The model-based clustering (probabilistic approach) is an alternative approach in which the model fits to the distribution of the data (the common ones are Gaussian mixture model and latent class analysis).

3.4.1 K-means Clustering

Throughout this part of the thesis, k-means clustering technique was used as a clustering algorithm to split observations into distinctive subgroups. It is a multivariate heuristic clustering algorithm that uses to detect distinctive subgroups in

heterogeneous population. The algorithm is originally developed by MacQueen in 1967 [105], and thereafter is modified considerably by a number of researchers. It categorizes the observations into non-overlapping K subgroups, in which each observation is assigned to the subgroup whose mean (centroid) is closest; and based on that categorization, a new group mean is determined. These steps continue until no single observation changes subgroup (see figure 3.5 on page 65 for graphical demonstration of the algorithm). In other word, the classification is based upon the placing of observations into more or less homogeneous clusters, which attempts to have more in common within a group than between subgroups (minimizing within group variation and maximizing between groups variation) [106]. The algorithm steps are summarized below.

K-means Clustering Algorithm

- **Step 1:** Randomly chooses K observation as representative of initial centroids (clusters).
- **Step 2:** Measures the distance between each observation and each centroid (e.g. using Euclidean distance), and assigns the observation into the centroid (cluster) in which it has the shortest distance.
- **Step 3:** Computes K new centroids by averaging the observations in each cluster.
- **Step 4:** Repeats steps 2 and 3 till non of the updated centroids differ from the previous iteration, or no observation changes cluster.
- **Step 5:** Returns the current set of clusters.

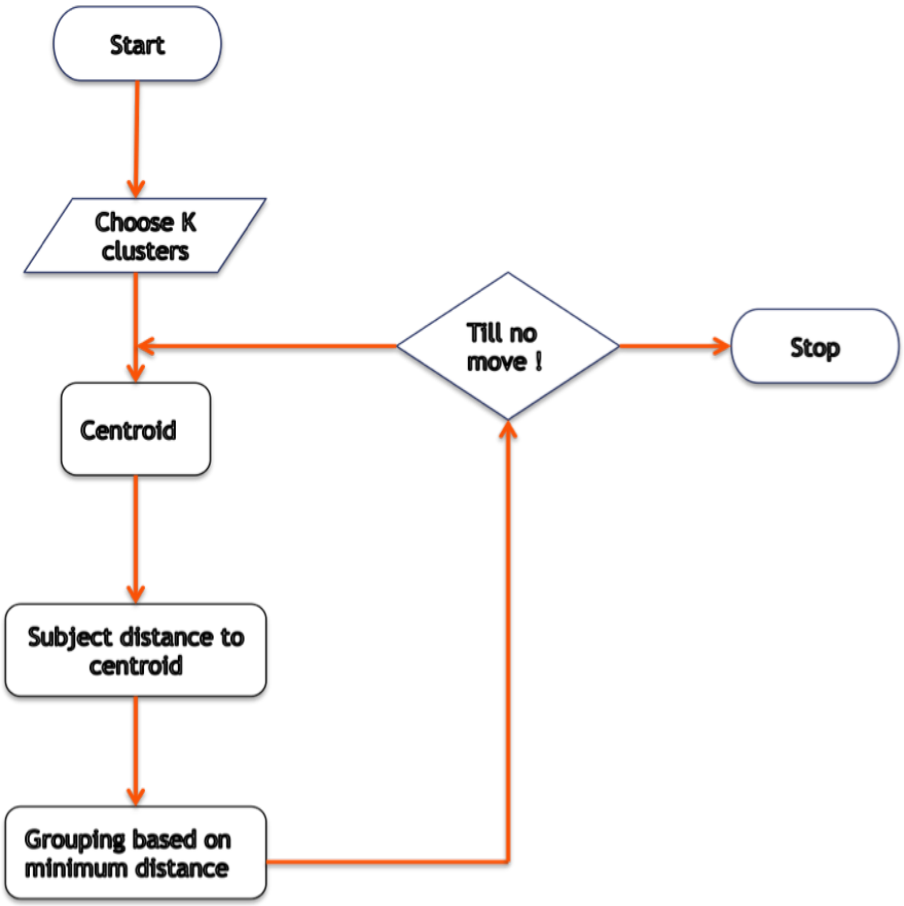


Figure 3.5: Graphical demonstration of k-means clustering algorithm

The k-means clustering algorithm is a robust, computationally feasible and relatively easy to understand and implement compared to the alternative approaches. It was the 3^{rd} most algorithm that has been used in the last decade, in studies which were applied statistical and machine learning clustering/classification techniques [107]. The algorithm handles very large samples, and is not prone to model over-fitting (when the number of parameters are greater than the number of observations). Its convergence is guaranteed [108] unlike to the model-based clustering techniques. In addition, it is available in most open and commercial statistical softwares, and any researcher could replicate this work without being a software expert.

Distance measure

In the k-means clustering algorithm, Squared Euclidean Distance was used to measure the relative (resultant) distance between the observations. It is formulated as follows:

$$D_{ij} = \sqrt{\sum_{k=1}^{p}(Y_{ik} - Y_{jk})^2} \tag{3.5}$$

Where D_{ij} is the Squared Euclidean distance between individual (observation) i with variable values $Y_{i1}, Y_{i2}, ..., Y_{ik}$ and individual j with variable values Y_{j1}, Y_{j2}, ..., Y_{jk}.

3.5 Proposed Input Variables to the Clustering Algorithm

The twenty artificial variables were presented into several formats, in the hope that the right approach (input variables to the clustering algorithm) will be identified to glean the right information (clusters) exist in the simulated data. We started with the standard approach, which uses all the observed (simulated) variables as input into clustering algorithm. In addition, data reduction using factor analysis (principal factor with varimax rotation) was implemented, and several representative of the simulated variables were extracted, such as latent variables (factor scores), highest-loading variables, and variables which have high-error terms, and were independently used as input variables into the k-means clustering algorithm. Finally, the corresponding clusters were identified using each input variables, and their performance were assessed by comparing the identified clusters against the known simulated clusters.

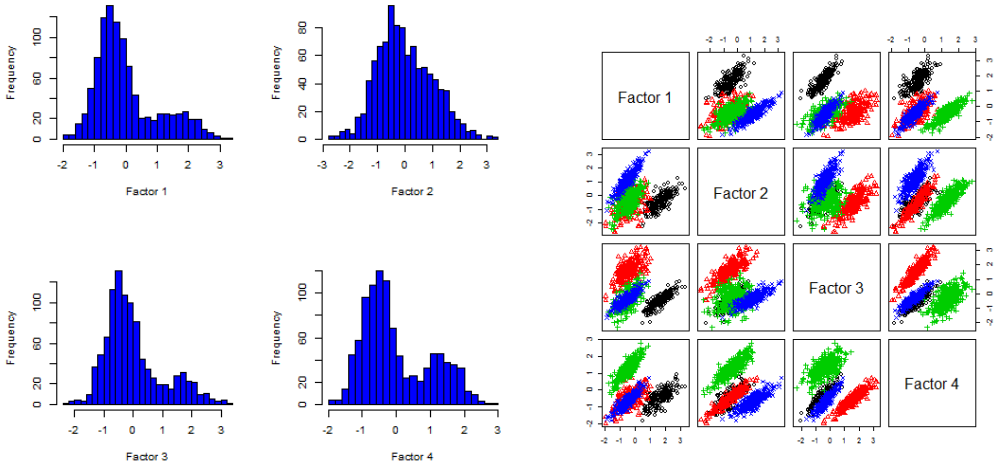
3.5.1 Clustering on All Observed Variables

In the standard approach (i.e. using all the twenty simulated variables as input into k-means clustering), four clusters were identified. About 62.8% of the 1000 observations were assigned into the true known clusters, and the rest 372 observations were assigned to the wrong clusters (misclassified).

3.5.2 Clustering on Factor Scores (Latent Variables)

The dimension of the twenty simulated variables was reduced into four independent factors using factor analysis, and the retained factors were rotated using varimax rotation (see section 3.3 for details). Then four corresponding independent factor scores (latent variables) were generated for each observation. These four factor scores were used as input variables into the k-means clustering algorithm, and four clusters were identified with 0% misclassification (i.e. all the observations were assigned correctly into the known simulated clusters).

The factor scores distributions (histograms) are displayed in figure 3.6(a) and the scatterplot across the subgroups (each cluster is represented by distinct colour) is depicted in figure 3.6(b) on page 67.



(a) (b)

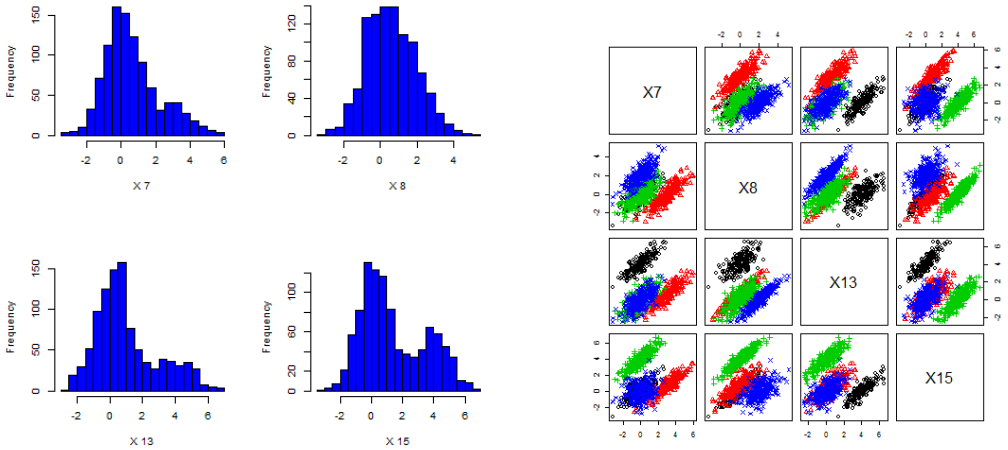
Figure 3.6: Factor scores: (a) distributions and (b) scatterplot across the subgroups

In the histogram above, we observe that majority of the factor scores have a bimodal distribution in which a factor can split the observations into at least two distinctive subgroups. In addition, it is clearly observed in the above scatterplot that each factor score has specific contribution in splitting the clusters. For example, factors 1 and 2 split the black, green and blue clusters very well, but not the green from the red cluster. In addition, factors 1 and 3 split the black, blue and red clusters, but not the blue from the green cluster. Furthermore, factors 1 and 4 split the black, blue and green clusters, but not the red from the blue clusters. So the combination of all the four factors able to split the four clusters extremely well without any misclassification.

3.5.3 Clustering on the Highest-loading Observed Variables

After factor analysis was implemented to the twenty simulated variables, and the four retained factors were rotated with varimax rotation. Then four corresponding observed variables which have the highest-loading in each factor were identified (see table 3.2 on page 62). Thus, X_{13} has the highest-loading in factor 1, X_8 in factor 2, X_7 in factor 3 and X_{15} in factor 4. Their distributions and scatterplots are depicted in figure 3.7 on page 68.

Then these four highest-loading observed variables (X_7, X_8, X_{13} and X_{15}) were used as input variables into k-means clustering, and four optimal clusters were identified and 99.9% of the observations were correctly assigned to the true known clusters.



(a) (b)

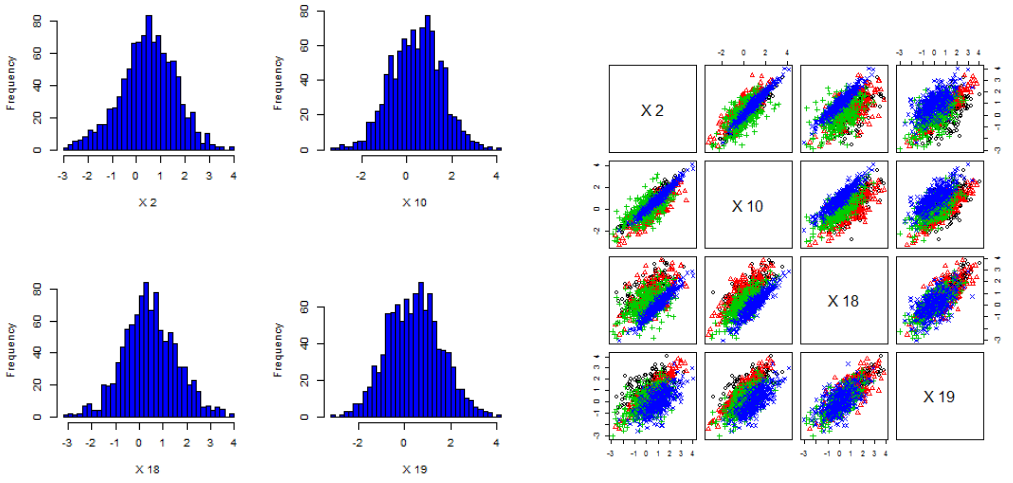
Figure 3.7: Highest-loading variables: (a) distributions and (b) scatterplot across the subgroups

In the histogram above, majority of the highest-loading variables have a bimodal distribution which suggested a single observed variable could able to split the observations into at least two distinctive subgroups. In addition, in the above scatterplot the four highest-loading variables were displayed across the four clusters, and each cluster is represented by distinct colour. The patterns of the four observed highest-loading variables across the clusters are very similar to patterns observed with respect to the factor scores across the clusters. Variables X_7 and X_8 split the red, blue and the green clusters very well, but not the black from the green cluster.

In addition, X_7 and X_{13} split the red, blue and black clusters, but not the blue from the green cluster. Furthermore, variables X_7 and X_{15} separate the black, the blue and the green clusters, but not the blue from the red cluster. Thus, the combination of all the four highest-loading variables (X_7, X_8, X_{13} & X_{15}) able to classify the entire clusters very well.

3.5.4 Clustering on Observed Variables with Highest Error-terms

For illustration purpose after factor analysis was performed, four observed variables which have relatively highest error-terms (unexplained variance by the retained factors) were identified, and reported in table 3.2 on page 62. These are X_2, X_{10}, X_{18} and X_{19}, and were used as input variables into k-means clustering algorithm. As consequence, only 45.6% of the observations were classified correctly to the right clusters. The distributions and scatterplots of these variables are depicted in figure 3.8 on page 69.



(a) (b)

Figure 3.8: Variables with highest error-terms (a) distributions and (b) scatterplot across the subgroups

In the histogram above, we observe that these variables which have high error-terms have very smooth bell-shaped normal distribution, in which do not show any pattern of slitting the clusters into distinctive subgroups. In addition, we observed in the above scatterplot that the clusters are entirely overlapped with respect to these four observed variables which have high error-terms (X_2, X_{10}, X_{18} and X_{19}),

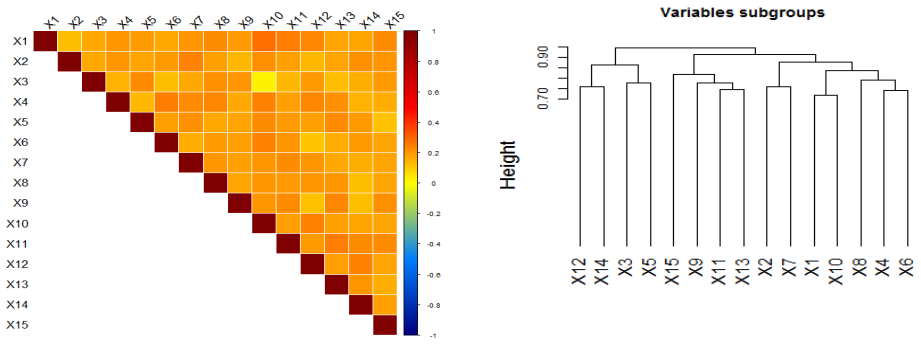
unlike to the pattern observed with respect to factor scores (latent variables), figure 3.6, and highest-loading observed variables, figure 3.7.

3.5.5 Noisy Observed Variables in Factor Analysis

To assess the influence of noisy variables in factor analysis, five random variables that do not have clustering information (which are also uncorrelated within themselves and with the original 20 simulated variables) were simulated. Then factor analysis was performed again to all the 25 simulated variables, and the optimal number of factors were retained (data not shown). As a consequence, these particularly noisy variables appeared to have high error-terms than their corresponding communalities (i.e. communality < 0.5). That means that these variables are not represented well by these retained factors. Therefore, factor analysis can be used as a screening approach for variable selection prior to generating the factor scores for cluster analysis, by removing these variables which have high error-terms than their corresponding communalities.

3.5.6 Application of Factor Analysis on Uncorrelated Observed Variables

In this section, a new data with 1000 observations which have measurements on 15 uncorrelated variables (correlation < 0.28) were simulated to investigated the robustness of factor analysis for this type of data. Prior to factor analysis, correlations of the variables are graphically presented as heatmap in figure 3.9 (a), and the variables subgroups were also assessed and illustrated in figure 3.9 (b) on page 71 .



(a) (b)

Figure 3.9: Simulated variables: (a) correlation matrix and (b) subgroups. Heatmap colors: dark-red represented for strong positive correlation; dark-blue for strong negative correlation; yellow for no correlation; light-red and light-blue for weak positive and negative correlation, respectively.

As illustrated in the heatmap above (figure 3.9 (a)), the correlations among the variables are very weak. In addition, the variables did not create any clear subgroups based on their correlation matrix (figure 3.9 (b)). This observation is in contrast to the pattern observed in the highly correlated variables which displayed in figures 3.1 (a) and (b) on page 54.

In addition, factor analysis (with varimax rotation) was applied to this data, and reduced to low dimensional factors. The factor loadings of the first four factors are displayed in table 3.3 on page 72.

Table 3.3: Varimax rotated factor loadings of the simulated data

Variable	Factor1	Factor2	Factor3	Factor4	C^1	U^2
X_1	0.26	0.15	0.19	0.32	0.23	0.77
X_2	0.31	0.21	0.18	0.1	0.19	0.81
X_3	0.07	0.43	0.15	0.14	0.23	0.77
X_4	0.42	0.18	0.09	0.18	0.24	0.76
X_5	0.17	0.2	0.27	0.23	0.2	0.8
X_6	0.39	0.12	0.21	0.07	0.21	0.79
X_7	0.29	0.24	0.16	0.2	0.2	0.8
X_8	0.31	0.28	0.11	0.17	0.22	0.78
X_9	0.23	0.22	0.31	0.08	0.2	0.8
X_{10}	0.43	0.00	0.18	0.27	0.29	0.71
X_{11}	0.24	0.18	0.33	0.2	0.24	0.76
X_{12}	0.21	0.2	0.09	0.39	0.25	0.75
X_{13}	0.23	0.16	0.31	0.18	0.2	0.8
X_{14}	0.17	0.16	0.25	0.26	0.19	0.81
X_{15}	0.21	0.25	0.22	0.16	0.18	0.82
Eigenvalue	1.16	0.71	0.71	0.69		

C^1 = Proportion of total variation accounted for by the common factors (common variance)
U^2 = Proportion of total variation not accounted by the common factors (unique variance)

As shown in the table above, the proportion of the shared variance which explained by the retained four factors is very small. For example, X_1 shared variance explained by the factors is only 0.23 (or interms of percentage is 23 %) and the rest 77% remained unexplained (as unique errors). This means that the variables do not have a noticeable shared variance. This shows that most of the information exist in the entire observed variables is not well represented by the retained factors (although more factors were retained than commonly extracted based on "screeplot" or "eigenvalue above one" criteria). For this type of data, it is highly unlikely that common latent factors would exist that could capture reasonably well the internal patterns of the observed variables. Thus, there is a real danger here in which researchers need to be aware, in which not to use factor analysis for variable reduction and extraction of the internal patterns when the observed variables are uncorrelated (or weakly correlated). In such situation, it is more likely to end up with latent factors which do not capture (represent) well most of the information exist in the

data, and its consequence could be quite dangerous.

3.5.7 Application of Factor and Cluster Analyses in Gene Expression

As we observed in the above simulation study, using factors scores as input into k-means clustering algorithm able to identify all the simulated clusters without any misclassification. Thus, for generalization, we applied these approach to gene expression data (which have similar patterns as the cytokines), to identify the existing clusters. The gene expression data, with repeated measurement, was generated by KY Yeung and colleagues, and were originally published here [109]. Several researchers had used these data to assess the performance of their new methods.

In this analysis, we used only the baseline measurements in which 20 columns (gene expression) and 400 rows (observations) with six known clusters. First, factor analysis (principal factor with varimax rotation) was preformed to this array data and reduced to three low dimensional factors, and subsequently estimated the corresponding factor scores (latent variables) for each observation and were used to identify the optimal clusters. As a consequence, all the six existing clusters were identified with 0% misclassification. In contrast, using all the 20 observed variables six clusters were identified with 18.5% misclassification. As the dimension of the dataset is relatively large, it is not displayed in this analysis but readers can refer to the original paper [109] for details.

3.5.8 Summary

In this chapter, the performance of several formats of observed variables as input into clustering algorithm were assessed on artificial data. The artificial data were simulated in the same format as the correlated real asthma and COPD sputum cytokines but with known class membership (see the sputum cytokines and simulated explanatory analysis for details in chapters 2 and 3, respectively). As a consequence, for this type of data using "factor scores" (derived from factor analysis), rather than the standard approach (using the actual measured variables as input into clustering algorithm) able to identify the true (optimal) clusters existing in the dataset. In addition, this approach was implemented in gene expression data (which are strongly correlated as the sputum cytokines), and able to identify all the existing clusters without any misclassification.

In addition, which worth mentioning that, using the highest-loading variables as input into clustering, appeared to perform very well in identifying the true clusters in the simulation study. However, this approach is not always guaranteed as sometimes multiple variables may have very close loadings in the same factor, and those loadings may not be as close as to the optimal (i.e. one). Thus, choosing only one representative variable for each factor may loss some information by underrepresenting the entire information exist within that specific factor. For example, it would be a dangerous practice to use the highest-loading variable as a representative of a factor in a situation where the variable's loading is relatively small (e.g. less than 0.7). Therefore, such technical issues need to be accounted when using only the highest-loading variables for further analysis such as cluster analysis. However, it is still worthy to identify variables which have the highest loadings (after varimax rotation) as they may give a general direction/suggestion which variables have substantial contribution in splitting the clusters, and fewer variables may be required for future validation of the clusters (such as assigning new observations into the existing clusters).

Furthermore, factor analysis can be used as screening technique for noisy variables in post analysis like cluster analysis, by removing variables in which their error terms are greater than their communalities (explained variance by the retained factors). However, as it has been demonstrated in this study, factor analysis is not an

appropriate approach to reduce the dimensionality and extract the internal patterns of uncorrelated variables. Thus, to cluster data which have no strong correlation among the variables, it is better to use all the observed variables or other alternative approaches instead of factor-scores which are derived from factor analysis.

In conclusion, incorporating variable selection into clustering algorithm needs many analytical decisions. For instance, decision should be made what type of variables to use as input into clustering algorithm, and whether dimensionality reduction (using factor analysis) for the extraction of the underlying structure is needed. This simulation study demonstrated how to assess those information (e.g. the underlying structure of the observed variables) and incorporate into the clustering algorithm. In the correlated artificial data, using factor scores as input into clustering performed best in identifying the simulated clusters. However, readers should be cautioned that by no means that we are claiming this approach as the optimal ones for any type of data. Despite the fact that clustering using the standard approach (using full-set of observed variables as input into clustering) could be easy to implement and might reveal the optimal clusters, but this approach may only work better in a situation where there is no hidden structure (no strong correlation) between the variables.

Chapter 4

Modeling Asthma and COPD Biological Heterogeneity at Stable State

4.1 Objectives

The objective of this chapter is to model the biological heterogeneity of asthma and COPD using the combination of factor and cluster analyses (unsupervised statistical techniques) using sputum cytokines (biological mediators) to identify the common and distinctive biological subgroups of both diseases at stable state.

4.2 Statistical Methods

A range of sputum mediators (cytokines) were recoded from asthmatic and COPD subjects at stable state. In the descriptive analysis, which was reported in chapter 2, there are considerable overlap between asthma and COPD with respect to sputum cytokines. In addition, the mediators appeared to correlated strongly and internal patterns/structures among the cytokines were observed (see figure 2.5). Furthermore, based on their correlation the mediators were partitioned into several subgroups (see figure 2.5(b) on page 49 for details). As it has been demonstrated in the simulation study (chapter 3) for this type of data using a two stage approach (factor scores as input into clustering algorithms) is the best method in identifying the optimal clusters. Therefore, this approach (factor and cluster analyses) was hypothesized to model the biological heterogeneity (using sputum cytokines) of asthma and COPD in order to identify the common and distinctive subgroups of the diseases. Factor analysis will be used to capture the profiles (internal patterns) of

the cytokines, and the k-means clustering to classify the individuals into distinctive subgroups who have similar biological profiles.

4.2.1 Application of Factor and Cluster Analyses in Asthma and COPD Study

Since the cytokines were negatively skewed, their natural logarithmic format were used for subsequent analysis. First, unsupervised multivariate modeling using factor analysis (principal factor with orthogonal varimax rotation) was performed on the cytokines, and a set of low-dimensional independent factors was obtained. However, prior to that data screening was performed using univariate descriptive statistics to assess the accuracy of the input observed variables to the factor analysis algorithm. Such as linearity and outliers were checked using matrix scatter plot (pairwise plots) [data not shown], but since outlier in cytokines have clinical meaning [110], they weren't excluded, but totally minimised by transforming to natural logarithm. In addition, variables were standardised to minimize the bias in weighting which may result from different ranges.

The optimal factors were retained on the basis of screeplot (factors above the break in the curve) and eigenvalue above one [102]. Mediators that have high collinearity [111] were excluded from factor analysis to avoid multicollinearity. As CXCL-10 and CXCL-11 were highly correlated, and CXCL-11's shared variance was better explained by the retained factors than was CXCL-10's variance, CXCL-10 was excluded from the model. Similarly, because IL-10 levels in more than one-third of asthmatic subjects were below the limit of detection, but the concentrations were not different between asthma and COPD then to avoid bias toward one disease, it was excluded from the model. No similar bias was observed for other mediators.

Subjects who did not have a complete record of the cytokine panel were excluded from the factor and cluster analysis. Factor scores were calculated for each subject using standardized values of the cytokines, inverse of the correlation matrix of the cytokines and factor loadings after varimax rotation (see the formula in equation 3.2 on page 57 for details). These scores represent the subjects predicted value for each factor and retain the relationship between factors, and were used as input into k-means clustering algorithm. Squared Euclidean distance (formulated in equation 3.5

on page 66) was used as a measure of similarity in the k-means clustering algorithm. The optimal number of clusters was chosen based on screeplot (clusters above the break in the curve) by plotting within cluster sum of the squares against a series of sequential number of clusters [112], and the pattern of the variables subgroups observed in figure 2.5(b) on page 49. In addition, it was assessed on the bases of how the clusters look on their clinical and biological implications and interpretability.

The logic behind the screeplot ("Elbow") method (for chosen the optimal clusters) is that, the within cluster sum of squares (WCSS) were calculated for each possible (k) clusters and drawn against a series of sequential number of clusters. The WCSS decrease as the number of clusters increases, and elbow point (inflation) on a Scree plot suggests where to cut off the possible number of clusters, and the number of clusters above that point is an appropriate number of groups [106].

All statistical analyses were performed using R version 3.2.1 [113], and STATA/SE version 13 [114]. The patterns of the demographic, clinical and biological characteristics (which described in chapter 2) were assessed further across the identified biological subgroups, in which normally distributed data were presented as mean (standard error of the mean), and log-transformed data as geometric mean with corresponding 95% confidence interval. The χ^2 test or Fisher exact test was used to compare proportions, and 1-way ANOVA was used to compare means across multiple groups. Non-normal data were presented as median with first and third quartiles, and Kruskal-Wallis test was used to compare subgroups. A p-value (two-sided) less than 0.05 was considered as statistically significant. The biological clustering results were interpreted with particular emphasis on biological profiles, demographic characteristics and their clinical implications.

4.2.2 Application of Linear Discriminant Analysis in Asthma and COPD Study

Linear discriminant analysis was performed to predict the identified biological clusters from factor scores using the actual measured cytokines. This approach was used to verify how well the clusters can be partitioned based on the measured cytokines, and subsequently were used to identify the individual cytokine contribution in discriminating the clusters. Then discriminant functions for each cytokine were calculated,

and corresponding discriminant scores (one less the number of clusters) for each subject was estimated (using the discriminant function or loadings of a cytokine and the original cytokine values for each subject) and were used to illustrate the subjects biological clusters graphically.

4.2.3 Linear Discriminant Analysis

Linear discriminant analysis is a supervised statistical technique of classification, which emphasizes the prediction of existing subgroup membership using new variables. It also uses for validation of the existing clusters/subgroups using a new dataset. More about its application for validation is described and discussed in the subsequent chapter.

Linear discriminant analysis is equivalent to posterior probability and is formulated as follows:

$$p(\pi_i/x) = \frac{\pi_i \, f(x/\pi_i)}{\sum_{i=1}^{G} \pi_i \, f(x/\pi_i)} \tag{4.1}$$

Where, G is the number of groups, and π_i is the prior probability for group i; $f(x/\pi_i)$ is group specific probability density function; and $p(\pi_i/x)$ is the posterior probability of group i given observation x.

The group specific probability density function can be written as follows:

$$f(x/\pi_i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\bigg(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i) \bigg) \tag{4.2}$$

Where, vector μ_i is group specific mean, and Σ is the pooled variance-covariance matrix which is common for all groups.

The parameters (mean and variance) can be estimated empirically from the sample. Then, subject can be assigned to the group in which he/she has the highest posterior probability.

4.3 Results

4.3.1 Asthma and COPD Biological Factors at Stable State

Among the 161 (asthma=86 and COPD=75) eligible subjects, 146 subjects with sixteen complete cytokines measurements were used to build factor analysis. The factor loadings matrix (after orthogonal varimax rotation) is displayed in table, 4.1 on page 81. The sixteen observed cytokines are appearing to be well represented and related to four independent unobservable (latent) factors. All factors were internally consistent and well represented the variables (cytokines), and the lowest communalities was 0.51. For example, the retained four factors explained 88%, 78%, and 75% of the shared variances of IL-1β, IL-5 and IL-6, respectively.

The loadings on F_1 are relatively higher for proinflammatory mediators (e.g. IL-1β, IL-6R, IL-8 and TNFα), but very small for the T_H2 derived cytokines (e.g. IL-5, IL-13, CCL-17 and CCL-26). The loadings on F_2 are very small for the T_H1 derived (e.g. CCL-5 and CXCL-11) and proinflammatory cytokines but relatively higher for these T_H2 derived cytokines. Thus, F_2 could be interpreted as a best representative of these T_H2 derived cytokines. The results from factor analysis supports the patterns observed in the visual inspection of the variables subgroups in figure 2.5(b) on page 49 in chapter 2, in which variables that grouped together, appeared to load in the same factor (table 4.1 on page 81).

Table 4.1: Varimax rotated factor loadings of sputum mediators at stable state

Variable	Factor1	Factor2	Factor3	Factor4	C^1	U^2
IL-1β	0.94	-0.04	0.00	-0.05	0.88	0.12
IL-5	0.16	0.84	0.21	0.10	0.78	0.22
IL-6	0.28	-0.07	0.81	0.02	0.75	0.25
IL-6R	0.73	0.23	0.12	0.21	0.65	0.35
IL-8	0.73	0.15	0.51	0.06	0.82	0.18
IL-13	0.08	0.74	-0.10	0.07	0.56	0.44
CCL-2	0.15	0.10	0.70	0.10	0.54	0.46
CCL-3	0.49	0.36	0.56	-0.05	0.69	0.31
CCL-4	0.42	0.36	0.59	0.00	0.66	0.34
CCL-5	0.65	0.27	-0.04	0.54	0.79	0.21
CCL-13	-0.14	0.43	0.64	0.18	0.65	0.35
CCL-17	-0.04	0.74	0.40	0.13	0.72	0.28
CCL-26	0.06	0.75	-0.04	0.24	0.63	0.37
CXCL-11	0.04	0.25	0.14	0.74	0.64	0.36
TNFα	0.88	-0.05	0.23	0.02	0.82	0.18
VEGF	0.57	0.11	0.35	0.29	0.54	0.46
Eigenvalue	4.05	3.05	2.92	1.12		

C^1 = Proportion of total variation accounted for by the common factors (common variance)
U^2 = Proportion of total variation not accounted by the common factors (unique variance)

4.3.2 Asthma and COPD Biological Clusters at Stable State

Using the combination of factor and cluster analyses (two stage approach), three distinct clinically relevant biological clusters (subgroups) of asthma and COPD subjects were identified that could not be determined using the existing guidelines; in which 58, 47 and 41 subjects were classified into cluster 1, 2 and 3, respectively. These clusters fairly represent Th-1, Th-2 and proinflammatory (PI), and PI dominant subgroups as determined by their sputum cytokine expression profiles, respectively. The clinical characteristics and the cytokines profiles across the subgroups are depicted in table 4.2 on page 82, and in table 4.6 on page 84, respectively.

Table 4.2: Statistical summaries of demographic and clinical characteristics across the three identified biological clusters at stable state that represent the differences and similarities between the subgroups

Variable	Cluster 1	Cluster 2	Cluster 3	P-value C1 vs. C2	P-value C1 vs. C3	P-value C2 vs. C3
Male [n (%)]	32 (55.2)	26 (55.3)	28 (68.3)	0.99	0.19	0.21
Current or Ex-smokers [n (%)]	22 (37.9)	29 (61.7)	40 (97.6)	0.015	<0.0001	<0.0001
Pack -year history	6.6 (3.8 - 11.6)	11.0(6.1 - 19.9)	40.3 (33.6 - 48.3)	0.13	<0.0001	<0.0001
Age (years)$^{+}$	55 (1.5)	60 (2.1)	67 (1.8)	0.038	<0.0001	0.008
Duration of Disease (years)	23 (17.2 - 29.6)	9 (5.8 - 12.8)	6 (4.0 - 7.9)	<0.0001	<0.0001	0.12
BMI (kg/m2)$^{+}$	30.2 (1.0)	28.8 (1.0)	25.3 (0.7)	0.28	<0.0001	0.005
Exacerbation number of steroids$^{\delta}$	3 (0.3)	3 (0.3)	4 (0.4)	0.31	0.002	0.047
Prednisolone dose use [n (%)]	34 (58.6)	19 (40.4)	6 (14.6)	0.06	<0.0001	0.001
Daily Prednisolone dose (mg)	10 (10 - 15)	7.5 (5 - 10)	5 (5 - 7.5)	0.008	0.006	0.5
Daily ICS dose (mcg/day)*[a]	1800 (1000 - 2000)	1000 (800 - 2000)	1000 (800 - 2000)	0.024	0.28	0.4
Pre FEV1 (L)$^{+}$	2.19 (0.1)	1.74 (0.1)	1.3 (0.1)	0.002	<0.0001	0.006
Pre FEV1/FVC ratio (%)$^{+}$	69.0 (1.9)	58.5 (2.3)	49.7 (2.4)	<0.0001	<0.0001	0.011
Pre FEV1 Predicted (%)$^{+}$	77.0 (2.7)	59.9 (3.7)	47.0 (3)	<0.0001	<0.0001	0.01
Post FEV1 (L)$^{+}$	2.35 (0.1)	1.88 (0.13)	1.37 (0.09)	0.005	<0.0001	0.003
Post FEV1 Predicted (%)+	81.7 (2.7)	63.9 (3.9)	49.1 (3)	<0.0001	<0.0001	0.005
Sputum Eosinophil count (%)	3.9 (2.4 - 6.4)	0.7 (0.5 - 0.9)	2.0 (1.25 - 3.17)	<0.0001	0.039	<0.0001
Sputum Neutrophil count (%)$^{+}$	58.8 (3.1)	77.18 (3)	59.1 (3.1)	<0.0001	0.95	<0.0001
Sputum Macrophage count (%)	16.6 (12.5 - 21.9)	12.2 (9.2 - 16.1)	25.7 (21.24 - 31.07)	0.1	0.026	<0.0001
TCC (x106 cells/g sputum)	1.31 (1.0 - 1.8)	4.6 (3.3 - 6.4)	1.8 (1.3 - 2.6)	<0.0001	0.15	<0.0001
Blood Eosinophil x109/L	0.24 (0.18 - 0.32)	0.25 (0.2 - 0.32)	0.21 (0.16 - 0.27)	0.73	0.54	0.28
Blood Neutrophil x109/L^{+}	5.74 (0.3)	5.82 (0.3)	5.77 (0.4)	0.85	0.94	0.92
Bacterial colonization (n[%])	8 (13.8)	26 (55.3)	9 (21.9)	<0.0001	0.29	0.001
VAS-cough (mm)$^{+}$	30 (3.0)	48 (4.0)	36 (4.4)	0.001	0.24	0.052
VAS-dyspnoea (mm)$^{+}$	31 (3.5)	46 (3.3)	46 (4.5)	0.003	0.006	0.93

Definition of abbreviations: VAS= Visual Analogue Score; BMI= Body Mass Index; ICS= Inhaled Corticosteroid; Daily Prednisolone dose = Daily Maintenance Prednisolone dose; FEV_1 = Forced Expiratory Volume in the First Second; FVC=Forced Vital Capacity; TCC=Total sputum cell count; C=cluster; Cluster 1= (Asthma=55; COPD=3); Cluster 2 = (Asthma=28; COPD=19); Cluster 3= (Asthma=2; COPD=39); CFU= colony forming units. Data presented as geometric mean (95% CI) unless stated;$^{+}$Mean (standard error of mean (SEM)); *median (1^{st} and 3^{rd} quartiles); Dose for only those subjects prescribed daily prednisolone; Pack-year history of current and ex-smokers; [a]beclomethasonedipropionate equivalent; $^{\delta}$= Total number of times a patient exacerbated and took high dose of steroids for at least three days in the last 12 months

Table 4.4: Statistical summaries of sputum mediators across the three identified biological clusters at stable state that represent the differences and similarities between the subgroups

Variable	Cluster 1	Cluster 2	Cluster 3	P-value C1 vs. C2	P-value C1 vs. C3	P-value C2 vs. C3
IL-1β (pg/ml)	39.5 (30.8 - 50.8)	379.5 (257.3 - 559.8)	23.5 (17.2 - 32.2)	<0.0001	0.025	<0.0001
IL-5 (pg/ml)	2.6 (1.6 - 4.2)	2.2 (1.4 - 3.4)	1.4 (0.9 - 2.2)	0.56	0.083	0.22
IL-6 (pg/ml)	21.3 (15 - 30.4)	271.4 (192.2 - 383.3)	486.2 (327.7 - 721.4)	<0.0001	<0.0001	0.031
IL-6R (pg/ml)	163.2 (126.0 - 211.6)	433.4 (344.2 - 545.6)	112.4 (88.6 - 142.6)	<0.0001	0.04	<0.0001
IL-8 (pg/ml)	1658 (1205 - 2280)	10884 (8709 - 13603)	3059 (2209 - 4236)	<0.0001	0.005	<0.0001
IL-10 (pg/ml)	0.33 (0.25 - 0.45)	5.5 (3.5 - 8.7)	0.34 (0.2 - 0.6)	<0.0001	0.89	<0.0001
IL-13 (pg/ml)	10.4 (7.7 - 14.0)	4.8 (3.8 - 6.2)	3.5 (2.4 - 5.2)	0.001	<0.0001	0.18
CCL-2 (pg/ml)	209.8 (168.3 - 261.5)	495.4 (378.1 - 649.1)	764.5 (538.8 - 1084.7)	<0.0001	<0.0001	0.055
CCL-3 (pg/ml)	20.2 (14.9 - 27.4)	97.4 (71.6 - 132.6)	47.9 (35.7 - 64.1)	<0.0001	<0.0001	0.002
CCL-4 (pg/ml)	237.8 (147.1 - 384.3)	1138.3 (847.8 - 1528.3)	807 (614.1 - 1060.5)	<0.0001	<0.0001	0.1
CCL-5 (pg/ml)	5.6 (4.5 - 7.0)	14.9 (11.1 - 20.1)	2.2 (1.8 - 2.8)	<0.0001	<0.0001	<0.0001
CCL-13 (pg/ml)	18.1 (12.9 - 25.5)	18.9 (13.6 - 26.2)	43.2 (32.6 - 57.2)	0.86	<0.0001	<0.0001
CCL-17 (pg/ml)	27 (19.2 - 37.9)	20.5 (14.0 - 30.0)	30.8 (21.5 - 44.2)	0.28	0.61	0.13
CCL-26 (pg/ml)	12.4 (7.8 - 19.9)	5.0 (3.4 - 7.5)	2.9 (1.9 - 4.6)	0.004	<0.0001	0.081
CXCL-10 (pg/ml)	418.7 (286.9 - 611.1)	860.1 (534.1 - 1384.9)	381.8 (262.5 - 555.3)	0.014	0.76	0.012
CXCL-11 (pg/ml)	34.1 (22.8 - 51.0)	42.5 (20.1 - 89.6)	19.2 (12.3 - 30.0)	0.56	0.15	0.089
TNFα (pg/ml)	1.4 (1. 1 - 1.9)	29.9 (19.5 - 45.9)	1.7 (1.1 - 2.5)	<0.0001	0.62	<0.0001
VEGF (pg/ml)	1020 (858 - 1213)	2199 (1871 - 2584)	1237 (1040 - 1471)	<0.0001	0.12	<0.0001

Definition of abbreviations: C= cluster; Cluster 1= (Asthma=55; COPD=3); Cluster 2 = (Asthma=28; COPD=19); Cluster 3= (Asthma=2; COPD=39). Data presented as geometric mean with corresponding 95% confidence interval (CI)

Table 4.6: Statistical summaries of sputum mediators across the three identified biological clusters at stable state that represent the differences and similarities between the subgroups

Variable	Cluster 1	Cluster 2	Cluster 3	P-value C1 vs. C2	P-value C1 vs. C3	P-value C2 vs. C3
IL-1β (pg/ml)	39.5 (30.8 - 50.8)	379.5 (257.3 - 559.8)	23.5 (17.2 - 32.2)	<0.0001	0.025	<0.0001
IL-5 (pg/ml)	2.6 (1.6 - 4.2)	2.2 (1.4 - 3.4)	1.4 (0.9 - 2.2)	0.56	0.083	0.22
IL-6 (pg/ml)	21.3 (15 - 30.4)	271.4 (192.2 - 383.3)	486.2 (327.7 - 721.4)	<0.0001	<0.0001	0.031
IL-6R (pg/ml)	163.2 (126.0 - 211.6)	433.4 (344.2 - 545.6)	112.4 (88.6 - 142.6)	<0.0001	0.04	<0.0001
IL-8 (pg/ml)	1658 (1205 - 2280)	10884 (8709 - 13603)	3059 (2209 - 4236)	<0.0001	0.005	<0.0001
IL-10 (pg/ml)	0.33 (0.25 - 0.45)	5.5 (3.5 - 8.7)	0.34 (0.2 - 0.6)	<0.0001	0.89	<0.0001
IL-13 (pg/ml)	10.4 (7.7 - 14.0)	4.8 (3.8 - 6.2)	3.5 (2.4 - 5.2)	0.001	<0.0001	0.18
CCL-2 (pg/ml)	209.8 (168.3 - 261.5)	495.4 (378.1 - 649.1)	764.5 (538.8 - 1084.7)	<0.0001	<0.0001	0.055
CCL-3 (pg/ml)	20.2 (14.9 - 27.4)	97.4 (71.6 - 132.6)	47.9 (35.7 - 64.1)	<0.0001	<0.0001	0.002
CCL-4 (pg/ml)	237.8 (147.1 - 384.3)	1138.3 (847.8 - 1528.3)	807 (614.1 - 1060.5)	<0.0001	<0.0001	0.1
CCL-5 (pg/ml)	5.6 (4.5 - 7.0)	14.9 (11.1 - 20.1)	2.2 (1.8 - 2.8)	<0.0001	<0.0001	<0.0001
CCL-13 (pg/ml)	18.1 (12.9 - 25.5)	18.9 (13.6 - 26.2)	43.2 (32.6 - 57.2)	0.86	<0.0001	<0.0001
CCL-17 (pg/ml)	27 (19.2 - 37.9)	20.5 (14.0 - 30.0)	30.8 (21.5 - 44.2)	0.28	0.61	0.13
CCL-26 (pg/ml)	12.4 (7.8 - 19.9)	5.0 (3.4 - 7.5)	2.9 (1.9 - 4.6)	0.004	<0.0001	0.081
CXCL-10 (pg/ml)	418.7 (286.9 - 611.1)	860.1 (534.1 - 1384.9)	381.8 (262.5 - 555.3)	0.014	0.76	0.012
CXCL-11 (pg/ml)	34.1 (22.8 - 51.0)	42.5 (20.1 - 89.6)	19.2 (12.3 - 30.0)	0.56	0.15	0.089
TNFα (pg/ml)	1.4 (1. 1 - 1.9)	29.9 (19.5 - 45.9)	1.7 (1.1 - 2.5)	<0.0001	0.62	<0.0001
VEGF (pg/ml)	1020 (858 - 1213)	2199 (1871 - 2584)	1237 (1040 - 1471)	<0.0001	0.12	<0.0001

Definition of abbreviations: C= cluster; Cluster 1= (Asthma=55; COPD=3); Cluster 2 = (Asthma=28; COPD=19); Cluster 3= (Asthma=2; COPD=39). Data presented as geometric mean with corresponding 95% confidence interval (CI)

4.3.3 Linear Discriminant Analysis Results

The measured cytokines were used to predict the three biological clusters (which were identified using factor scores as input into k-means clustering) using linear discriminant analysis. Subsequently, two (one less the number of clusters) discriminant functions (scores) for each subject were extracted. Thereafter, based on these scores, the subjects were presented graphically across the identified clusters in figure 4.1 on page 85. As we observed in the scatterplot, the clusters are well separated, in which the first discriminant function (x-axis) separates clusters 1 and 3 very well, but does not separate clusters 1 and 2, and clusters 2 and 3. The second discriminant function (orthogonal to the first) achieves reasonably well in separating clusters 1 and 2, and clusters 2 and 3 on the basis of associations not used in the first discriminant function. Therefore, the first and second discriminant functions together were used to represent the clusters and achieved extremely well in discriminating the three biological clusters/subgroups.

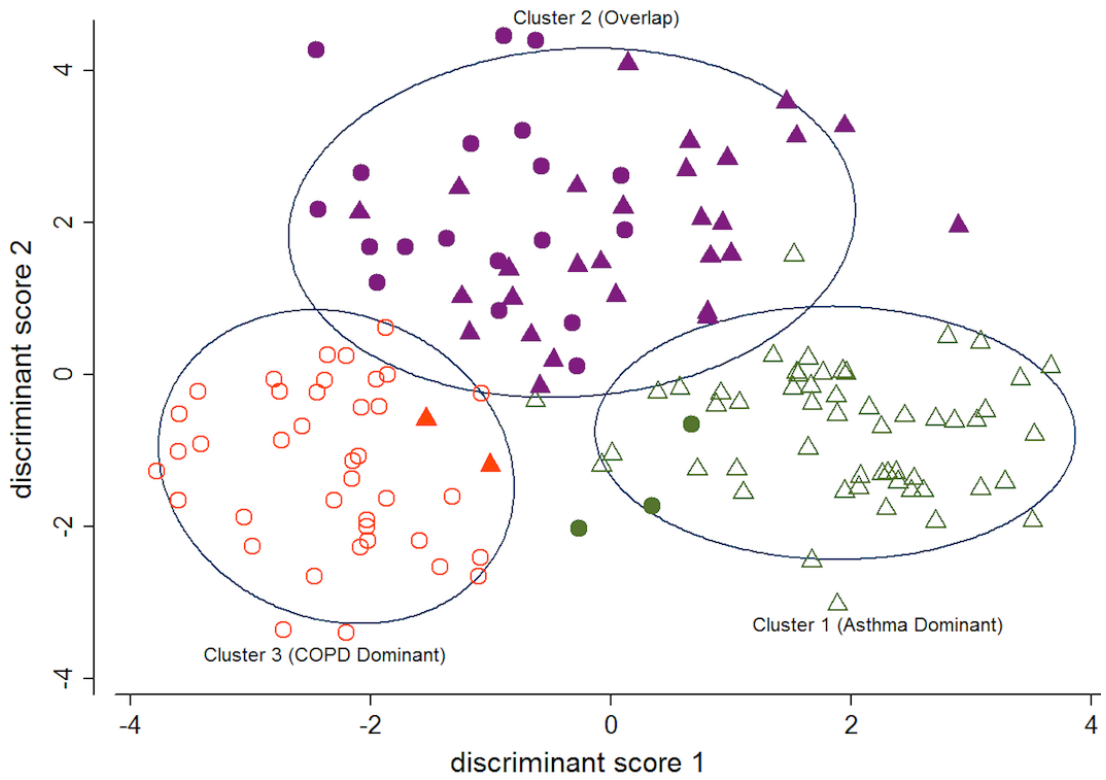


Figure 4.1: The three identified biological clusters presented using the subjects discriminant scores. Hollow triangle indicates eosinophilic asthma dominant (95% asthma, n=58); bold triangle and bold circle, neutrophilic asthma and COPD (overlap) dominant (59.6% asthma, n=47); hollow circle, COPD dominant (95% COPD, n=41); bold triangle, overlapped asthma; bold circle, overlapped COPD.

In addition, the overall patterns of the cytokines (using z-scores of the mediators) across the three clusters are illustrated graphically in figure 4.2 on page 86.

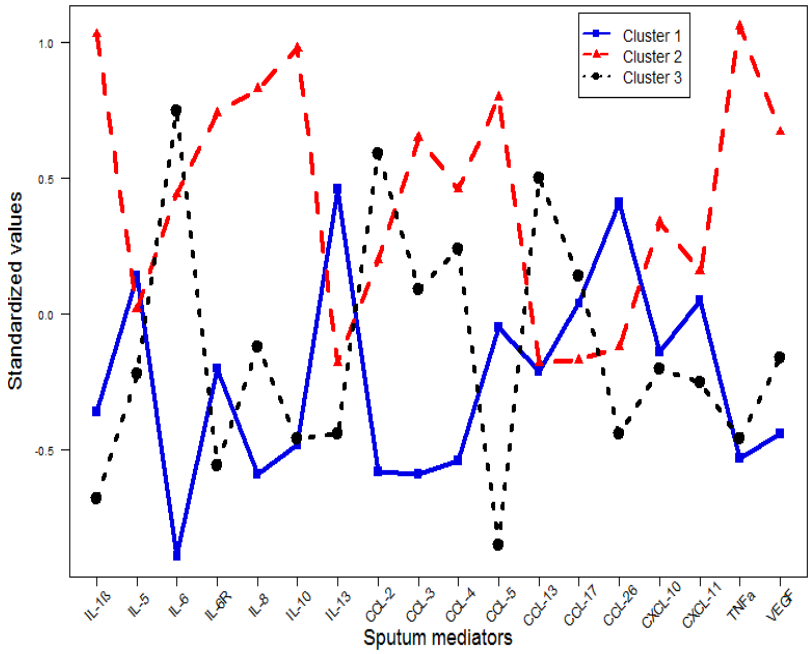


Figure 4.2: Patterns of sputum mediators across the identified clusters

Cluster 1

Cluster 1 consisted about 40% of the subjects included in this analysis (n=58), 95% were asthmatic, 55% were men and 38% were current or ex-smokers with average age of 55 years. In this group the mean (SEM) were 69% (1.9) and 81.7% (2.7) for FEV_1/FVC ratio and post FEV_1 percentage predicted, respectively (see table 4.2 on page 82 for details). Further stratification of cluster 1 by disease severity (airflow obstruction) based on the lung-function spirometry measurements (FEV_1/FVC ratio and post FEV_1 percentage predicted) showed that the subjects had 24.1% mild, 27.6% moderate, 5.2% severe and 0% very severe airflow obstruction.

In addition, the mean (95%CI) sputum eosinophil cell-counts was 3.9 (2.4 – 6.4)%, and the mean (SEM) sputum neutrophil cell-counts was 58.8% (3.1) (table 4.2 on page 82); in which 67% of subjects having a eosinophilia (differential sputum eosinophils cell-count $> 3\%$) and 48% a neutrophilia (differential sputum neutrophils cell-count $> 61\%$). Further stratification of cluster 1 by sputum cell counts showed that the subjects were 40% pure eosinophilic, 21% pure neutrophilic, 27% mixed granulocytic and 12% paucigranulocytic.

Furthermore, subjects in this cluster have elevated sputum T_H2 derived mediators (IL-5, IL-13 and CCL-26) compared to the other groups (see table 4.6 on page 84 and figure 4.2 on page 86).

Cluster 2

Cluster 2 consisted of an overlap of asthma and COPD (asthma=28, COPD=19), in which the subjects were 55% men, 62% were current or ex-smokers with 11 average pack–year history. This subgroup characterized by 58.5% (2.3) mean (SEM) FEV_1/FVC ratio and 63.9% (3.9) mean (SEM) post FEV_1 percentage predicted. Further stratification of this cluster by disease severity based on the lung-function spirometry measurements showed that 15.2% of the subjects had mild, 28.3% moderate, 19.6% severe, and 10.9% very severe airflow obstruction.

In addition, subjects in this cluster have a lower level of sputum percentage eosinophil in sputum {0.7% (95% CI: 0.5% – 0.9%; p-value <0.0001)} and higher percentage sputum neutrophil {77.2% (SEM=3), p-value <0.0001)} compared to cluster 1. In addition, in this cluster only 11% of asthmatic and 5% of COPD subjects had a sputum eosinophilia, but 75% of asthmatics and 95% of COPD had a sputum neutrophilia. Further stratification of this cluster with respect to sputum eosinophils and neutrophils showed that the subjects were 0% pure eosinophilic, 74% pure neutrophilic, 9% mixed granulocytic and 17% paucigranulocytic.

In general, subjects in this cluster characterized by high levels in total cell-counts, VAS-cough, VAS-dyspnoea, bacterial colonisation and sputum neutrophils (all p-values $< 0{\cdot}05$), but lower in sputum eosinophil and FEV_1/FVC ratio, pre-FEV_1 predicted, post-FEV_1 predicted, compared to subjects in cluster 1 (all p-values $< 0{\cdot}05$). In addition, several proinflammatory and T_H1 derived mediators (such as IL-1β, IL-6, IL-6R, IL-8, IL-10, CCL-2, CCL-3, CCL-4, CCL-5, CXCL-10, TNFα and VEGF) were significantly elevated in this cluster compared to cluster 1 (all p-value < 0.05).

Cluster 3

This cluster is a COPD predominantly group, in which 95% of the subjects were diagnosed as COPD patients, and about 68% were men and 98% were current or

ex-smokers with average 40 pack-year history. In addition, this subgroup characterized by 49.7% average FEV_1/FVC ratio (SEM=2.4) and 49.1% average post FEV_1 percentage predicted (SEM=3), which were significantly lower in this cluster compared to clusters 1 and 2 (all p-value <0.05). Further stratification of cluster 3 by disease severity on the basis of lung-function measurements revealed that 2.6% of the subjects had mild, 33.3% moderate, 33.3% severe, and 17.9% had very severe airflow obstruction.

In addition, subjects in this cluster have on average 2.0% sputum eosinophil (95% CI: 1.25% – 3.17%), and 59.1% sputum neutrophil (SEM=3.1). Further stratification of this cluster on cell-counts showed that the subjects were 21% pure eosinophilic, 28% pure neutrophilic, 23% mixed granulocytic and 28% paucigranulocytic.

Moreover, subjects in this cluster were substantially different from cluster 1 patients with respect to demographic and clinical characteristics (all p-values < 0·001), except in sputum neutrophil and total cell-counts (all p-values > 0·05). In addition, there are considerable differences between this cluster and cluster 2 in terms of demographic, clinical and lung function characteristics (all p-values < 0·05), but not in VAS-cough and VAS-dyspnoea (all p-values > 0·05) (see table 4.2 for details).

Furthermore, with respect to biological mediators, IL-6, IL-8, CCL-2, CCL-3, CCL-4 and CCL-13 are significantly higher, but IL-1β, IL-6R, IL-13, CCL-5 and CCL-26 are significantly lower in this cluster compared to cluster 1 (all p-value < 0.05). However, IL-6 and CCL-13 are significantly increase, but IL-1β, IL-6R, IL-8, IL-10, CCL-3, CCL-5, CXCL-10, TNFα and VEGF are significantly decrease in this cluster compared to cluster 2 (all p-value < 0.05) (see table 4.6 for details).

4.4 Discussion

In this study, the biological heterogeneity of asthma and COPD were modeled jointly using the combination of factor and cluster analyses. Subsequently, identified three distinctive biological subgroups of the diseases with different proportion of overlap. These subgroups were asthma dominant (Cluster 1), asthma and COPD overlap group (Cluster 2), and COPD predominant subgroup (Cluster 3). The findings have further emphasized the complex heterogeneity of asthma and COPD and provided support for the "British" hypothesis of airway disease pathogenesis as we identified 2 clusters that were predominately either asthma or COPD with distinct cytokine profiles, while also supporting the "Dutch" hypothesis by identifying a third cluster of overlapping subjects from both disease groups with similar cytokine profiles.

Cluster 1 was asthma predominant with evidence of eosinophilic inflammation and increased T_H2 inflammatory mediators. Cluster 2 contained an asthma and COPD overlap group, with predominately neutrophilic airway inflammation and elevated levels of IL-1β and TNF-α in addition to being assigned the highest proportion of subjects with bacterial colonization. Cluster 3 was a COPD-predominant group with mixed granulocytic airway inflammation and high sputum IL-6 and CCL-13 levels.

The clusters we have identified have biological plausibility and they confirm and extend our current understanding of the diseases beyond previous comparisons of asthma versus COPD [92] or clustering approaches of cytokine profiles in asthma [115] or COPD [90]. In addition, the clusters might represent groups with possible stratified responses to specific anti-inflammatory treatment. Cluster 1 is consistent with the T_H2 predominant eosinophilic asthma paradigm. Indeed, this group was predominately asthmatic but importantly also included about 5% of subjects with COPD. It would seem likely that this group would respond to anti-T_H2 cytokine therapy such as anti–IL-5 [72, 73, 116, 117]. Whether subjects with COPD in this cluster would respond to anti-T_H2 cytokine therapy is currently under study here www.clinicaltrials.gov NCT01227278 [74].

Cluster 2 included an overlap of subjects with asthma and COPD. This group was predominately neutrophilic, consistent with previous observations [118], with

increased bacterial colonization. In this cluster, increased bacterial colonization was evident perhaps suggesting that in these subjects the neutrophilic inflammation is more likely a consequence of bacterial colonization rather than the primary abnormality; however, further studies are required for generalization.

Cluster 3 included mainly subjects with COPD in which bacterial colonization was observed in fewer subjects in spite of consistently elevated proinflammatory cytokines. Perhaps this group, in contrast to cluster 2, represents subjects in which the proinflammatory environment plays a more causal role in the disease expression rather than as a consequence of infection.

This study may bring new definitions that acknowledges the overlap and highlights the similarities and differences between the two diseases. In addition, it may bring attention to the potential contributions of cytokines in the classification asthma and COPD phenotypes, which might yield new insights that could benefit future efforts in diagnosis, prevention and more personalised intervention (treatment-specific anti-inflammatory therapies).

One possible limitation of this study is that only subjects with severe asthma and COPD who attended a secondary care setting were included, and thus might not be representative of a more generalized population. We acknowledge that our findings cannot be extrapolated to mild to moderate asthma or mild COPD but are confident that our populations are representative of our broader secondary care patient population. Further studies are required to include healthy controls, larger disease populations including a broader spectrum of subjects including those with mild disease from multicenters.

In conclusion, we found here that sputum inflammatory mediator profiling can determine distinct and overlapping groups of subjects with asthma and COPD. We identified an asthma-predominant cluster with eosinophilic inflammation and elevated T_H2 inflammatory mediators, a COPD-predominant group with elevated proinflammatory cytokines, and an asthma and COPD overlap group that clinically had chronic bronchitis, increased bacterial colonization, elevated sputum IL-1β and TNF-α levels, and a sputum neutrophilia. We predict that these groups might contribute to improved patient classification to enable a stratified medicine approach to airways disease.

Chapter 5

Asthma and COPD Validation at Stable State

5.1 Objectives

This chapter will focus on the validation of the stable biological subgroups that were identified in chapter 4, using new independent asthma and COPD study.

5.2 Introduction

Three distinctive asthma and COPD biological subgroups were identified at stable state (see chapter 4 for details). To validate the patterns of these subgroups, independent 166 severe asthma and 58 COPD subjects (all at stable state) were included in this study. In this validation study, a number of demographic and clinical characteristics such as gender, smoking status, age, height, weight, age of disease on-set, lung-function measurements and sputum cell-counts; and a range of sputum cytokines mediators were recorded.

5.3 Descriptive Analysis of Validation Study

The patterns of the demographic and clinical characteristics were assessed across asthma and COPD and depicted in table 5.1 on page 92.

Table 5.1: Statistical summaries of demographic and clinical characteristics across asthma and COPD in the validation study that represent the similarities and differences between the two diseases

	Asthma		COPD		
Variable	N	Mean (95% CI)	N	Mean (95% CI)	P-value
Male (%)	166	98 (59)	58	41 (70.7)	0.11
Current or Ex- smokers [n (%)]	166	56 (33.7)	58	58 (100)	< 0.0001
Pack -year history^{s}	45	8 (5.53 - 11.65)	56	41 (34.45 - 48.88)	< 0.0001
Age (years)$^{+}$	166	50.3 (1.07)	58	69.2 (1.24)	< 0.0001
Duration of Disease (years)	164	18.5 (16.09 - 21.31)	55	3.9 (3.17 - 4.73)	< 0.0001
BMI (kg/m2)$^{+}$	165	29.8 (0.52)	55	27.3 (0.82)	0.02
Exacerbation number of steroids$^{\delta}$	166	3.3 (0.24)	56	2.8 (0.35)	0.21
Prednisolone dose use [n (%)]	166	89 (53.6)	58	3 (5.2)	< 0.0001
Daily Prednisolone dose (mg)*	89	10 (7.5 - 15)	3	5 (5 – 7.5)	0.04
Daily ICS (mcg/day)^{*a}	166	1600 (1000 - 2000)	58	1000 (200 - 2000)	< 0.001
Pre FEV1/FVC ratio (%)$^{+}$	161	68.7 (1.01)	58	53.2 (1.76)	< 0.0001
Pre FEV1 Predicted (%)$^{+}$	161	70.6 (1.73)	55	56 (2.86)	< 0.0001
Sputum Neutrophil count (%)$^{+}$	148	62.6 (2.13)	58	66.7 (3.04)	0.3
Sputum Eosinophil count (%)	148	5.1 (4.2 - 6.2)	58	2.9 (2.21 - 3.83)	0.002
TCC ($x10^{6} cells/gsputum$)	151	1.4 (1.13 - 1.8)	56	3.8 (2.82 - 5.11)	< 0.0001

Definition of abbreviations: BMI= Body Mass Index; FEV_1=Forced Expiratory Volume in the First Second; FVC=Forced Vital Capacity; TCC=Total sputum cell count; Inhaled Corticosteroid dose=ICS; Prednisolone dose use = Maintenance prednisolone dose use. Data presented as geometric mean (95% CI) unless stated;$^{+}$Mean (standard error of mean (SEM)); *median (1st and 3rd quartile); Dose for only those subjects prescribed daily prednisolone; ^{s}Pack-year history of current and ex-smokers; ^{a}beclomethasonedipropionate equivalent; $^{\delta}$= Total number of times a patient exacerbated and took high dose of steroids for at least three days in the last 12 months.

As we observed in the above table, the two validation diseases have very distinctive demographic and clinical characteristics, which is pretty similar to the patterns observed in the test asthma and COPD study (see chapter 2 for details).

In addition, the patterns of sputum mediators were assessed across asthma and COPD in the validation study, and depicted in table 5.2 on page 93.

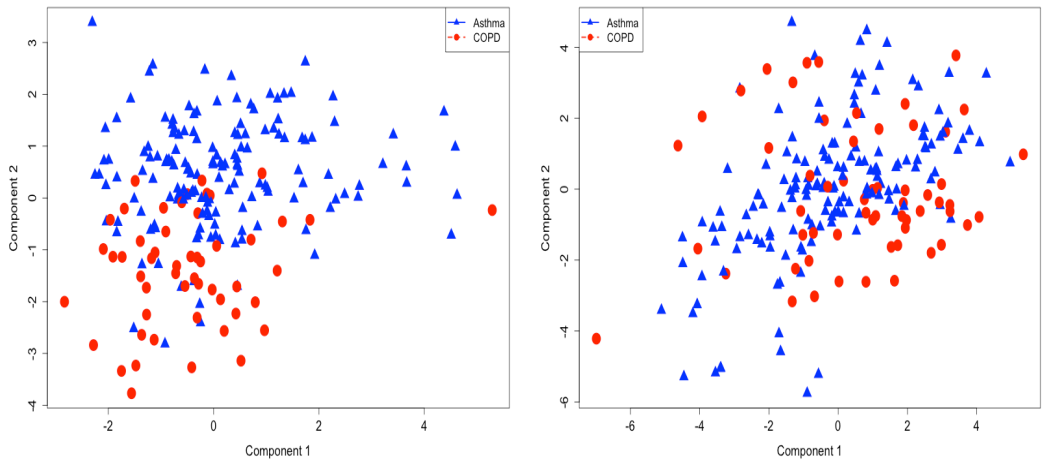
Table 5.2: Statistical summaries of sputum mediators across asthma and COPD in the validation study that represent the similarities and differences between the two diseases

	Asthma		COPD		
Variable	N	Mean (95% CI)	N	Mean (95% CI)	P-value
IL-1β	156	104.7 (79.9 - 137.1)	58	138.8 (86.4 - 223)	0.3
IL-5	165	1.3 (1.1 - 1.6)	58	1 (0.6 - 1.6)	0.21
IL-6	165	60.6 (47.1 - 77.8)	58	222.6 (141.6 - 349.9)	< 0.0001
IL-6R	156	294.8 (223.5 - 388.8)	58	149.1 (115.3 - 193)	0.006
IL-8	156	2683 (1869 - 3851	58	2418.5 (1811 - 3229)	0.74
IL-10	159	2.6 (2.2 - 3)	58	0.9 (0.5 - 1.5)	<0.0001
CCL-2	165	248.3 (213.9 - 288.3)	58	551.9 (414.5 - 734.8)	<0.0001
CCL-4	156	191.8 (140.5 - 261.7)	58	940 (597.8 - 1477.8)	<0.0001
CCL-5	165	7.4 (6.1 - 9)	58	5.6 (4.1 - 7.5)	0.15
CCL-13	159	16.6 (14.3 - 19.3)	58	33.5 (24.2 - 46.3)	<0.0001
CCL-17	165	22 (18.1 - 26.7)	58	30.1 (20.6 - 44.1)	0.12
CXCL-10	156	410.3 (301.7 - 557.9)	58	275.9 (186 - 409.2)	0.16
CXCL-11	165	22.3 (16.9 - 29.3)	58	14.1 (8.1 - 24.7)	0.12
TNFα	165	3.5 (2.7 - 4.5)	58	4.6 (2.5 - 8.2)	0.33

Data presented as geometric mean with correspoding 95% confidence interval (CI); unit of the cytokines is pg/ml

As shown in the table above, the two diseases have distinctive mediators, but majority of the mediators are similar across the two diseases (mediators with no significant difference across the two diseases).

Furthermore, to investigate the overall patterns of the characteristics across the validation asthma and COPD, principal component analysis (PCA) was performed separately for the combination of demographic and clinical parameters, and for sputum mediators (cytokines) and displayed graphically across the first two PCA scores in figure 5.1 (a) and (b) on page 94, respectively.



(a) (b)

Figure 5.1: Validation asthma and COPD study presented using the first two principal component scores: (a) Demographic and clinical characteristics; (b) Sputum mediators (cytokines).

As we observe in figure 5.1 above, the validation asthma and COPD subjects are quite distinctive with respect to the demographic and clinical characteristics; however, there is considerable overlap between the two diseases on their sputum mediators (cytokines). These patterns are very similar to the patterns observed in the test (original) study (see figure 2.2 on page 45 and figure 2.4 on page 48 in chapter 4, respectively, for details).

5.4 Validation using Linear Discriminant Analysis

Linear discriminant analysis (LDA) was proposed as a validation technique for this study. LDA (formulated in equations 4.1 and 4.2 on page 79) is a supervised statistical technique which can be used to predict the known subgroups/clusters using a new dataset. It implements by developing a linear discriminant function (classification model) from the original dataset for each subgroup (as formulated in equation 5.1 on page 95). Then the new dataset is plugged-in into each subgroup's classification model and corresponding discriminant score for each observation in each cluster is calculated, and the observation is assigned into the subgroup in which the individual has the highest discriminant score.

5.4.1 Linear Discriminant Functions

Linear discriminant analysis is formulated in equation 4.1 and 4.2 on page 79, chapter 4. Then linear discriminant function (LDF) is derived for each subgroup and formulated as follows:

$$LDF_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + log\,(\omega_k) \tag{5.1}$$

Where, $\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + log\,(\omega_k)$ is the constant, and $\Sigma^{-1}\mu_k$ are the coefficients in each group, in which Σ^{-1} is the inverse of the pooled variance-covariance matrix; and μ_k and ω_k are the mean and the proportion of observations in each group, respectively.

The parameters (ω_k, μ_k & Σ) can be estimated empirically from the sample as follows:

$$\begin{aligned} \omega_k &= \frac{N_k}{N} \\ \mu_k &= \frac{\sum_{i=1}^{N_k} x_i}{N_k} \\ \Sigma &= \sum_{k=1}^{K}(x_i - \mu_k)(x_i - \mu_k)^T/(N-K) \end{aligned} \tag{5.2}$$

Where, N_k is the number of samples in each group K, and $\sum_{k=1}^{K} N_k = N$ is the total sample size.

After some mathematical transformation, equation 5.1 (classification model) can be rewritten as follows:

$$D_{ij} = \beta_j + \beta_{j1}X_{1i} + \beta_{j2}X_{2i} + ... + \beta_{jk}X_{ki} + log(P_j) \tag{5.3}$$

Where D_{ij} is the discriminant score for subject i in group j; β_j is a constant for the j^{th} group, β_{jk} is the weight (coefficient) for variable k in group j; X_{ki} is the observed value of subject i on the k^{th} variable; $log(P_j)$ is a logarithmic scale of prior probability of group j membership.

5.4.2 Validation in Simulation Study

Although LDA was proposed to validate the identified subgroups of asthma and COPD using independent study, unfortunately the number of cytokines which were recorded in the validation studies were relatively smaller (i.e. 75%) compared to the cytokines that used in the identification (construction) of the original subgroups/clusters in the test study in chapter 4. Therefore, this issue has been approached by proposing that the the original subgroups need to be predicted using only cytokines that exist in the validation study. Thereafter, new classification model for each cluster using the existing cytokines will be built, and subsequently subjects will be assigned to the subgroup in which he/she has the highest discriminant score. However, no previous studies applied this approach for validation using fewer variables compared to the original variables that used in the identification of the subgroups. Therefore, prior to applying this approach to real asthma and COPD validation study, it will be applied to a simulation study to investigate its robustness/validity.

A new artificial data which has the same patterns, number of observations and variables as the one simulated in chapter 3 (simulation study) was simulated. Then 75% of these variables were randomly selected and matched with the original simulated variables (from chapter 3). The proportion of the selected new simulated variables are compatible with the proportion of the cytokines that exist in asthma and COPD validation study to the original cytokines which were used for the identification of the biological clusters of the diseases in chapter 4.

The simulated clusters (which were correctly identified using factor scores as input into k-means clustering in chapter 3) were predicted using these selected variables using linear discriminant analysis. A classification model for each cluster was developed, and the *betas* (classification functions) for each variable in each cluster with prior class probabilities are depicted in table 5.3 on page 97. Subsequently, the validation study was plugged-in into this model, and each observation was assigned to the subgroup in which the individual has the highest discriminant score.

Table 5.3: Coefficients (βs) and class proportion (prior probabilities) in each cluster in the test simulated study were used to predict class membership in the validation simulated study

Variable	Group 1 (β_{1k})	Group 2(β_{2k})	Group 3 (β_{3k})	Group 4 (β_{4k})
X_1	20.0	5.3	7.7	-5.0
X_2	31.9	5.4	-11.9	10.2
X_3	49.7	1.0	-13.8	7.9
X_4	1.7	0.1	0.2	9.6
X_6	-24.8	20.7	-11.5	13.5
X_8	3.4	-1.8	-31.1	32.6
X_{10}	-30.9	-4.1	7.8	1.8
X_{11}	-7.8	-10.1	31.1	-27.3
X_{12}	-15.0	-5.6	10.5	-18.3
X_{13}	100.8	19.2	-9.3	-20.0
X_{14}	-62.4	-12.3	1.5	9.4
X_{15}	-5.2	1.6	23.1	-10.0
X_{16}	2.7	2.2	11.1	-36.8
X_{17}	-48.1	-7.5	-5.8	41.0
X_{19}	-14.6	-15.0	-9.5	-4.8
Constant	-209.3	-24.3	-83.3	-68.5
Prior probability	0.25	0.25	0.25	0.25

Prior probabilities are equal across the subgroups in the test study.

Once the new simulated data was plugged-in into the above table 5.3, fortunately, all the observations were assigned into the right simulated subgroups although partial (75%) of the original simulated variables were used for observations assignment. Thus, as the approach appeared more robust in the simulation study, it was applied to asthma and COPD study to validate the biological clusters that were identified in chapter 4.

5.4.3 Validation in Asthma and COPD Study

As described above, the mediators (cytokines) that were measured in the validation study are relatively small compared to the ones that were recorded in the test (original) asthma and COPD study (which was used in the identification of the biological clusters). Therefore, the identified asthma and COPD clusters were predicted only using the mediators that exist in the validation study. The classification model for each cluster was developed, and the *betas* (classification functions) for each mediator in each cluster with prior class probabilities are depicted in table 5.4 on page 98. Finally, the discriminant score of each validation subject in each group

was calculated, and the subject was assigned into the subgroup in which he/she has the highest discriminant score.

Table 5.4: Coefficients (βs) and class proportions (prior probabilities) in each cluster in the test asthma and COPD study that were used to predict class membership in the validation study

Variables	Cluster 1 (β_{1k})	Cluster 2 (β_{2k})	Cluster 3 (β_{3k})
IL-1β	1.63	2.17	0.96
IL-5	-6.31	-6.5	-6.79
IL-6	-2.62	-1.17	0.14
IL-6R	6.28	5.58	4.72
IL-8	6.64	7.59	7.31
CCL-2	5.56	5.66	6.98
CCL-4	3.59	3.83	4.66
CCL-5	-1.36	-0.8	-2.83
CCL-13	-0.04	-0.49	-0.21
CCL-17	2.13	1.09	1.09
CXCL-11	0.48	0.51	0.29
TNFα	-5.77	-5.28	-6.57
Constant	-63.39	-75.75	-79.11
Prior Probability	0.33	0.33	0.33

Prior probabilities are equal across the subgroups in the test study.

The summary statistics of the clinical parameters and mediators of the validation study of asthma and COPD across the subgroups are presented in tables 5.5 and 5.6 on page 99, respectively.

Table 5.5: Statistical summaries of demographic and clinical characteristics across the validation subgroups (which represent the differences and similarities between the subgroups) that were predicted using linear discriminant analysis

Variable	Group 1	Group 2	Group 3	P-value G1 vs. G2	P-value G1 vs. G3	P-value G2 vs. G3
Male [n (%)]	65 (61.3)	40 (54.8)	25 (71.4)	0.38	0.28	0.1
Current or Ex-smokers	37 (34.9)	42 (57.5)	33 (94.3)	0.003	<0.0001	<0.0001
Pack-year history	13.2 (8.5 - 20.7)	16.0 (10.0 - 25.6)	39.6 (31.3 - 50.1)	0.52	<0.0001	0.002
Age (year)$^{+}$	53 (1.3)	56 (2.0)	66 (2.0)	0.15	<0.0001	0.002
Duration of Disease (year)	15 (12.9 - 18.7)	13 (9.8-16.8)	6 (4.2 - 8.5)	0.24	<0.0001	0.002
BMI (kg/m2)$^{+}$	30.2(0.7)	29.0 (0.8)	27.1 (0.9)	0.22	0.017	0.13
Prednisolone dose use [n (%)]	54 (50.9)	30 (41.1)	7 (20.0)	0.19	0.001	0.03
Daily Prednisolone dose (mg)	10 (10 - 15)	10 (5 - 15)	5 (5 - 10)	0.52	0.03	0.12
Daily ICS dose (mcg/day)*[a]	2000 (1000 - 2000)	1600 (1000 - 2000)	1000 (200 - 2000)	0.2	<0.001	0.008
Pre FEV_1/FVC ratio (%)$^{+}$	68.2 (1.4)	61.8 (1.9)	57.3 (1.8)	0.004	<0.0001	0.14
Pre FEV_1 Predicted (%)$^{+}$	67.6 (2.2)	64.7 (2.9)	65.5 (3.2)	0.4	0.65	0.85
Sputum Neutrophil count (%)$^{+}$	59.1(2.8)	72.3 (2.5)	58.4 (4.0)	0.001	0.88	0.003
Sputum Eosinophil count (%)	6.3 (4.9 - 8.2)	3.0 (2.4 - 3.8)	3.2 (2.3 - 4.4)	<0.0001	0.003	0.79
TCC (x106 cells/g sputum)	1.2 (0.9 - 1.5)	3.5 (2.5 - 4.8)	2.99 (2.07 - 4.3)	<0.0001	0.001	0.6

Definition of abbreviations: BMI= Body Mass Index; FEV_1=Forced Expiratory Volume in the First Second; FVC=Forced Vital Capacity; TCC=Total sputum cell count; G=Group; Group 1=(Asthma=94; COPD=12); Group 2= (Asthma=55; COPD=18); Group 3= (Asthma=7; COPD=28); CFU= colony forming units; ICS= Inhaled Corticosteroid dose; Prednisolone dose use = Maintenance prednisolone dose use. Data presented as geometric mean (95% CI) unless stated;$^{+}$Mean (standard error of mean (SEM)); *median (1^{st} and 3^{rd} quartiles); Dose for only those subjects prescribed daily prednisolone; Pack-year history of current and ex-smokers; [a]beclomethasonedipropionate equivalent.

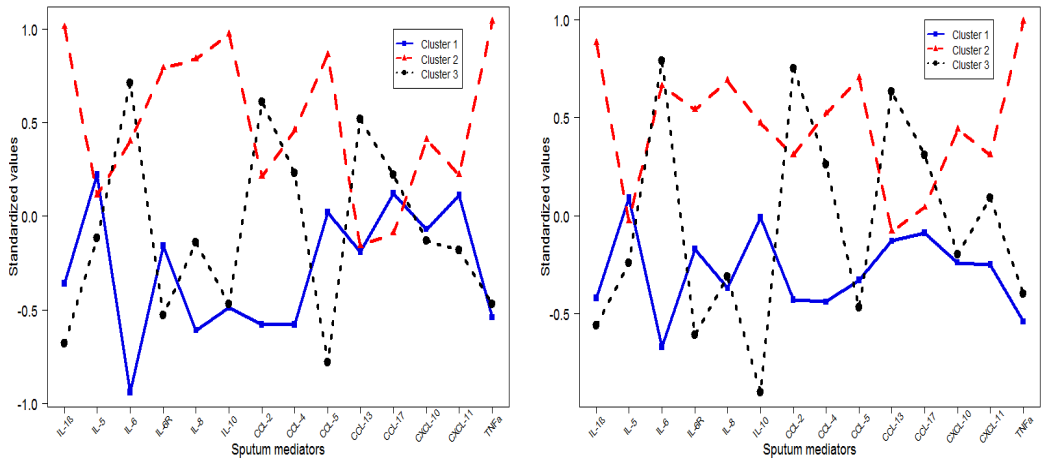
Table 5.6: Statistical summaries of sputum mediators across the validation subgroups (which represent the differences and similarities between the subgroups) that were predicted using linear discriminant analysis

Variable	Group 1	Group 2	Group 3	P-value G1 vs. G2	P-value G1 vs. G3	P-value G2 vs. G3
IL-1β (pg/ml)	54.1 (42.7 - 68.5)	526.6 (375.5 - 738.4)	42.4 (26 - 69.2)	<0.0001	0.36	<0.0001
IL-5 (pg/ml)	1.4 (1.1 - 1.8)	1.2 (0.8 - 1.7)	0.9 (0.5 - 1.5)	0.43	0.085	0.34
IL-6 (pg/ml)	26.1 (19.9 - 34.3)	273.3 (210.6 - 354.6)	344.2 (237.2 - 499.3)	<0.0001	<0.0001	0.32
IL-6R (pg/ml)	186.0 (135.5 - 255.3)	589.8 (476.6 - 729.8)	90.5 (50.3 - 162.8)	<0.0001	0.013	<0.0001
IL-8 (pg/ml)	1210 (822 - 1783)	10771 (8846 - 13115)	1387 (643 - 2989)	<0.0001	0.7	<0.0001
IL-10 (pg/ml)	1.9 (1.5 - 2.4)	3.9 (2.8 - 5.4)	0.5 (0.3 - 0.9)	0.001	<0.0001	<0.0001
CCL-2 (pg/ml)	191.9 (161.7 - 227.8)	425.6 (338.6 - 535.0)	680.1 (486.9 - 949.8)	<0.0001	<0.0001	0.025
CCL-4 (pg/ml)	119.6 (80.3 - 178.1)	848.1 (646.8 - 1111.9)	502.5 (248 - 1018.2)	<0.0001	<0.0001	0.1
CCL-5 (pg/ml)	4.5 (3.7 - 5.6)	16.8 (13.3 - 21.2)	3.8 (2.6 - 5.5)	<0.0001	0.39	<0.0001
CCL-13 (pg/ml)	17.4 (14.5 - 21.0)	18.4 (14.2 - 23.9)	40.0 (27.7 - 57.9)	0.73	<0.0001	0.001
CCL-17 (pg/ml)	21.3 (16.4 - 27.7)	25.2 (19.0 - 33.4)	36.0 (23.3 - 55.6)	0.4	0.043	0.17
CXCL-10 (pg/ml)	235.8 (167.7 - 331.7)	841.5 (566.5 - 1249.9)	254.1(140.7 - 458.8)	<0.0001	0.83	0.001
CXCL-11 (pg/ml)	12.4 (8.7 - 17.6)	36.1 (22.1 - 59.1)	23.6 (15.4 - 35.9)	<0.0001	0.08	0.28
TNFα (pg/ml)	1.4 (1.2 - 1.7)	23.3 (16.1 - 33.7)	1.8 (1.1 - 3.0)	<0.0001	0.32	<0.0001

Definition of abbreviations: G=Group. Data presented as geometric mean with corresponding 95% confidence interval; Group 1=(Asthma=94; COPD=12); Group 2= (Asthma=55; COPD=18); Group 3= (Asthma=7; COPD=28)

Three validation subgroups were identified using linear discriminant analysis, in which group 1 consists of 88% asthmatics and 12% COPD subjects; group 2 comprises 75% asthmatic and 25% COPD subjects; and group 3 consists of 20% asthmatic and 80% COPD subjects. The subjects in group 1 have elevated per-

centage sputum eosinophils; group 2 subjects have high level in sputum neutrophils, and in most of the proinflammatory mediators and T_H1 derived mediators. Whereas group 3 is non-neutrophilc COPD dominated group in which the subjects have high level of IL-6 and CCL-2. The overall patterns of the mediators across the test (original) clusters and the validation subgroups are depicted graphically in figure 5.2 on page 100.



(a) (b)

Figure 5.2: Patterns of mediators : (a) across test clusters and (b) across validation subgroups which validated using linear discriminant analysis

Overall, the clinical and mediators patterns observed in the validation study are very similar to the test study (see the validation tables 5.5 and 5.6, and the test tables 4.2 and 4.6; and the test and validation figures 5.2(a) and 5.2(b) for detailed comparisons).

5.5 Validation using IL-1β and Disease Status

Assignment of a new subject to a pre-existing subgroup (validation) using linear discriminant analysis is mathematically plausible. However, as it happened in this study, it needs a considerable number of variables (cytokines) to assign a single subject to the right subgroup as the original clusters were identified using a range of cytokines. This approach is too complicated to implement in clinic for subject assignments into relevant subgroup. Therefore, a simple alternative approach was proposed, which will be implemented based on the observed patterns of the clinical and biological characteristics across the test clusters. Then the validation subjects will be assigned to the subgroups using this approach, with the hope that not too much information is lost compared to the alternative linear discriminant analysis approach.

A clear pattern was observed in the original identified clusters, in which it was evident that cluster 1 (asthma dominated) and cluster 3 (COPD predominant) can be easily split based on the clinical characteristics of asthma and COPD, according to the GINA and GOLD guidelines. However, it is quite difficult to identify the overlap group (cluster 2) based on these guidelines. Therefore, to establish the cutoff value from the cytokines for cluster 2, a Classification and Regression Trees (CART) technique was performed, using RPART R package [119], one at a time to all these cytokines which have the highest discriminant function (from discriminant analysis) and highest factor loadings (from factor analysis).

Those cytokines that performed well in discriminating the overlapping cluster at a cutoff which found using CART, were compared using their percentage of correctly predicted values (sensitivity ratios). Then the best cutoff was established according to the higher total accuracy of the confusion matrix. Subsequently, the best determined cytokine cutoff (with the highest sensitivity ratio in discriminating the clusters), together with the disease classification (asthma or COPD), were applied to classify the validation study into three subgroups.

Thus, IL-1β at 130 pg/ml cutoff performed extremely well in discriminating the overlap group (cluster 2) from cluster 1 and 3 in the test study, and is depicted in

figure 5.3 on page 102. IL-1β cutoff with combination of disease status (asthma or COPD) were applied as a classifier to validate the identified clusters using independent asthma and COPD study.

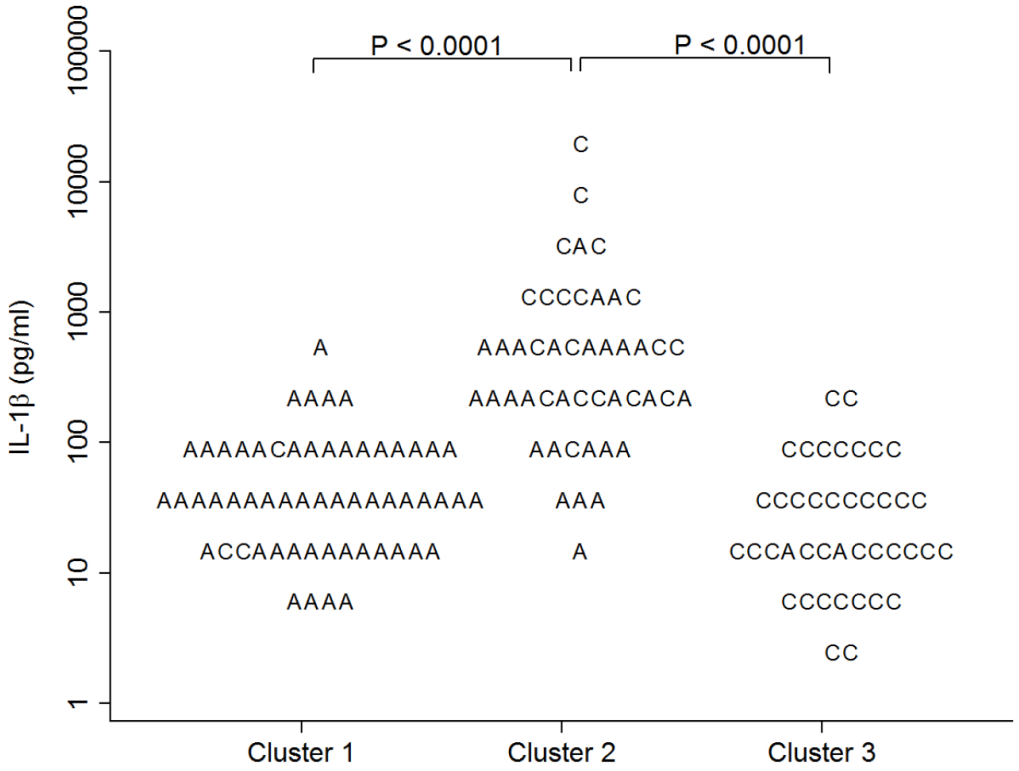


Figure 5.3: Absolute IL-1β concentrations on a log scale (base 10) across the 3 identified stable biological clusters. A= Asthma; C=COPD. P is the p-value for geometric mean difference between cluster 1 or cluster 3 versus cluster 2 (overlap).

Assuming asthma and COPD as two clinically distinctive diseases (according to the existing guidelines) and IL-1β above 130 pg/ml cutoff for the overlap, the validation subjects were assigned to three subgroups. For instance, if a subject was diagnosed as asthmatic and his/her IL-1β level is below 130 pg/ml he/she would be assigned to group 1; and if a subject was diagnosed with COPD and his/her IL-1β level is below 130 pg/ml was assigned to group 3; and irrespective of disease status (asthma or COPD) if a subject has IL-1β level above 130 pg/ml was assigned to group 2 (overlap cluster). Therefore, all the clinical characteristics and sputum mediators of the validation study were reported across the subgroups in tables 5.7 and 5.8, respectively, on page 103.

Table 5.7: Statistical summaries of demographic and clinical characteristics across the validation subgroups (which represent the differences and similarities between the subgroups) that were predicted using IL-1β and disease status (asthma or COPD)

Variable	Group 1	Group 2	Group 3	P-value G1 vs. G2	P-value G1 vs. G3	P-value G2 vs. G3
Male [n (%)]	66 (64.1)	51 (57.3)	22 (68.7)	0.34	0.63	0.26
Current or Ex-smokers	28 (27.2)	54 (60.7)	32 (100)	<0.0001	<0.0001	<0.0001
Pack-year history	9.5 (5.8 - 15.7)	17.0 (11.5 - 25.2)	41.5 (33.5 - 51.6)	0.06	<0.0001	0.001
Age (year)$^{+}$	49 (1.2)	57 (1.8)	68 (1.6)	<0.0001	<0.0001	0.001
Duration of Disease (year)	17 (14.1 - 20.7)	13 (10.6 - 16.6)	4 (2.7 - 4.7)	0.08	<0.0001	<0.0001
BMI (kg/m2)$^{+}$	29.9 (0.7)	29.0 (0.6)	27 (1.3)	0.33	0.03	0.13
Prednisolone dose use [n (%)]	57 (55.3)	35 (39.3)	2 (6.25)	0.027	<0.0001	<0.0001
Daily Prednisolone dose (mg)*	10 (8 - 15)	10 (5 - 10)	6.25 (5 - 7.5)	0.15	0.11	0.24
Daily ICS dose (mcg/day)*[a]	1600 (1000 - 2000)	2000 (1000 - 2000)	900 (200 - 2000)	0.52	0.006	0.008
Pre FEV1/FVC ratio (%)+	70.1 (1.2)	61.0 (1.7)	56.6 (2.4)	<0.0001	<0.0001	0.16
Pre FEV_1 Predicted (%)$^{+}$	71.2 (2.3)	64.1(2.5)	60.2 (3.6)	0.033	0.019	0.4
Sputum Neutrophil count (%)$^{+}$	59.0 (2.8)	70.4 (2.5)	59.2 (4.2)	0.003	0.97	0.023
Sputum Eosinophil count (%)	5.6 (4.3 - 7.3)	3.5 (2.8 - 4.4)	3.9 (2.5 - 5.9)	0.009	0.12	0.68
TCC (x106 cells/g sputum)	0.98 (0.7 - 1.3)	3.1 (2.3 - 4.1)	3.1 (2.2 - 4.5)	<0.0001	<0.0001	0.97

Definition of abbreviations: BMI= Body Mass Index; FEV_1=Forced Expiratory Volume in the First Second; FVC=Forced Vital Capacity; TCC=Total sputum cell count; G=Group; Group 1= (Asthma=103, COPD=0); Group 2 = (Asthma=63 and COPD=26); Group 3= (Asthma = 0 and COPD=32); CFU = colony forming units; Prednisolone dose use = Maintenance prednisolone dose use; ICS = Inhaled Corticosteroid. Data presented as geometric mean (95% CI) unless stated;$^{+}$Mean (standard error of mean (SEM)); *median (1st and 3rd quartile); Dose for only those subjects prescribed daily prednisolone; Pack-year history of current and ex-smokers; [a]beclomethasonedipropionate equivalent

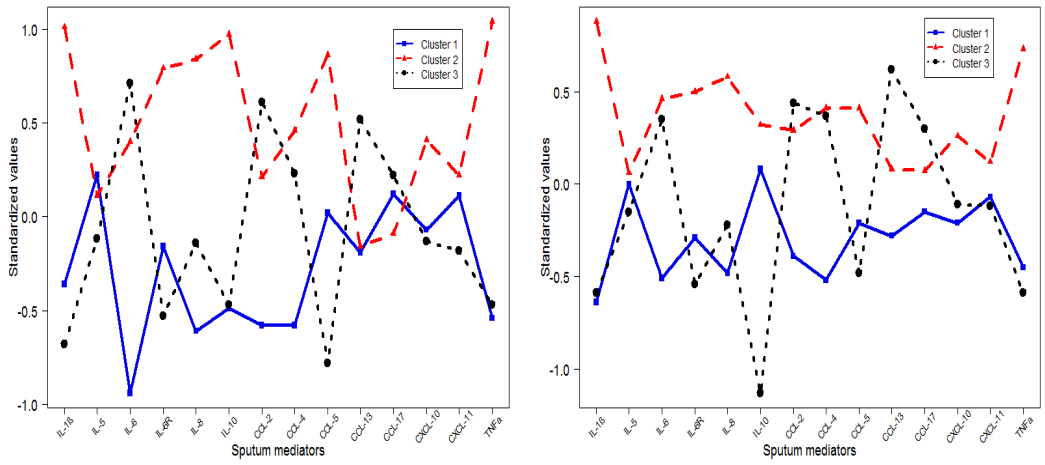
Table 5.8: Statistical summaries of sputum mediators across the validation subgroups (which represent the differences and similarities between the subgroups) that were predicted using $IL-1\beta$ cutoff and disease status (asthma or COPD)

Variable	Group 1	Group 2	Group 3	P-value G1 vs. G2	P-value G1 vs. G3	P-value G2 vs. G3
IL-1β (pg/ml)	37.0 (29.1 - 47)	527.1 (407.1 - 682.5)	40.0 (28.2 - 56.6)	<0.0001	0.75	<0.0001
IL-5 (pg/ml)	1.2 (1.0 - 1.5)	1.3 (1.0 - 1.9)	1.0 (0.6 - 1.7)	0.69	0.45	0.37
IL-6 (pg/ml)	34.7 (25.6 - 47)	190 (138.4 - 261)	157.7 (88.8 - 280.2)	<0.0001	<0.0001	0.56
IL-6R (pg/ml)	153 (102.7 - 228.6)	549.4 (454.9 - 663.6)	101.7 (74.1 - 139.5)	<0.0001	0.17	<0.0001
IL-8 (pg/ml)	975 (597 - 1592)	8609 (706 2 - 10496)	1646 (1041 - 2603)	<0.0001	0.15	<0.0001
IL-10 (pg/ml)	2.2 (1.8 - 2.6)	3.1 (2.2 - 4.3)	0.4 (0.2 - 0.6)	0.063	<0.0001	<0.0001
CCL-2 (pg/ml)	202 (166.8 - 244.9)	414.8 (336.4 - 511.6)	488.3 (340.0 - 701.4)	<0.0001	<0.0001	0.44
CCL-4 (pg/ml)	101.3 (65.3 - 157.3)	685.4 (497.8 - 943.8)	631.5 (382.8 - 1041.7)	<0.0001	<0.0001	0.79
CCL-5 (pg/ml)	5.2 (4.1 - 6.7)	11.7 (9.2 - 14.8)	3.7 (2.5 - 5.5)	<0.0001	0.16	<0.0001
CCL-13 (pg/ml)	14.7 (12.4 - 17.5)	22.0 (17.1 - 28.3)	39.4 (27.4 - 56.6)	0.01	<0.0001	0.017
CCL-17 (pg/ml)	19.5 (15.1 - 25)	26.1 (19.7 - 34.7)	35.7 (22.9 - 55.6)	0.13	0.025	0.26
CXCL-10 (pg/ml)	250.5 (163 - 384)	595.4 (420.6 - 842.9)	297.7 (194.2 - 456.6)	0.002	0.64	0.034
CXCL-11 (pg/ml)	17.3 (12.4 - 24.2)	25.0 (16.3 - 38.5)	15.8 (7.8 - 31.8)	0.18	0.81	0.28
TNFα (pg/ml)	1.6 (1.3 - 2.0)	14.4 (9.8 - 21.2)	1.3 (0.7 - 2.1)	<0.0001	0.39	<0.0001

Definition of abbreviations: G=Group; Group 1= (Asthma=103 and COPD=0); Group 2 = (Asthma=63 and COPD=26); Group 3= (Asthma = 0 and COPD=32). Data presented as geometric mean with corresponding 95% confidence interval

Thus, three validation subgroups were identified using IL-1β cut off and the disease status (asthma or COPD), in which group 1 consists of 100% asthmatics (n=103), group 2 consists 71% (n=63) asthmatic and 29% (n=26) COPD subjects,

and group 3 consists of 100% (n=32) COPD subjects. The subjects in group 1 have elevated percentage sputum eosinophils; group 2 subjects have high level in sputum neutrophils, and in most of the proinflammatory mediators and T_H1 derived mediators. Whereas group 3 is non-neutrophilc COPD dominated group with high level of IL-6 and CCL-2. The overall patterns of the mediators across the test (original) clusters and the validation subgroups are depicted graphically in figures 5.4(a) and 5.4(b) on page 100.



(a) (b)

Figure 5.4: Patterns of sputum mediators : (a) across test clusters and (b) across validation subgroups using IL-1β cutoff and disease status

The patterns of the clinical and sputum mediators in the validation subgroups (which identified using IL-1β and disease status) is very similar to the test clusters; (see the validation tables 5.7 and 5.8, and the test tables 4.2 and 4.6; and the test and validation figures 5.4(a) and 5.4(b) for details).

TNFα

In this study, although IL-1β at cutoff 130 pg/ml performed really well in discriminating the overlap group (cluster 2) from clusters 1 and 2, TNFα also performed similarly in discriminating the overlap group at cutoff 5 pg/ml (see figure 5.5 on page 105). Thus, TNFα could be an alternative marker to IL-1β for future subjects assignment to the identified biological subgroups.

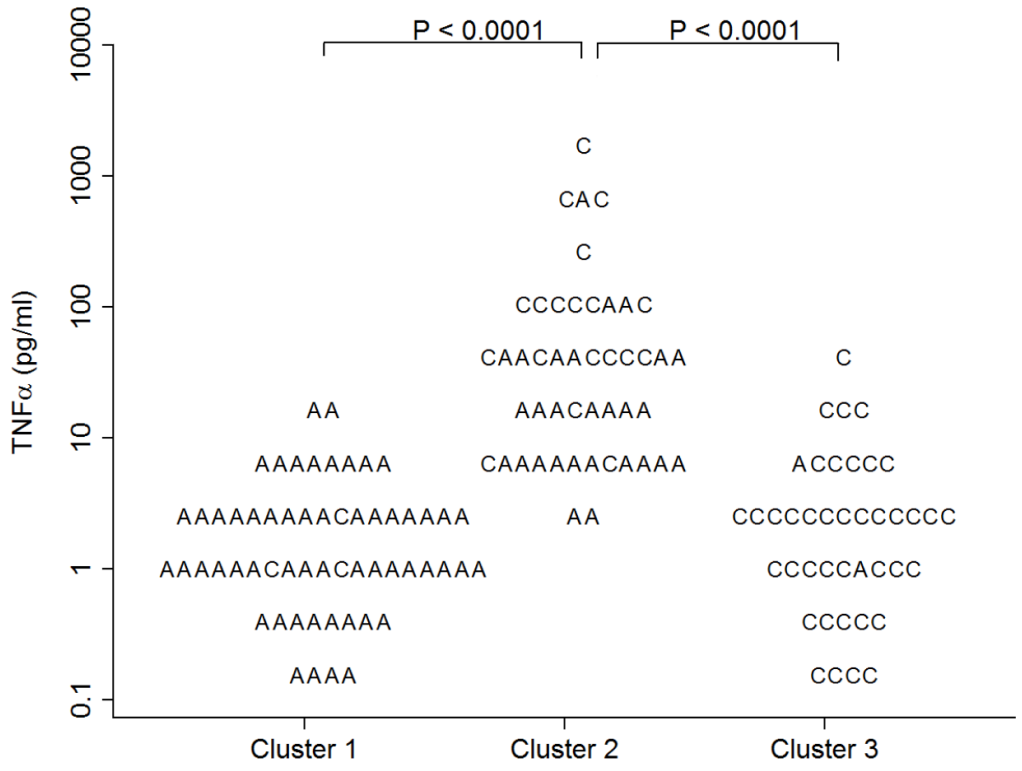


Figure 5.5: Absolute $TNF\alpha$ concentrations on a log scale (base 10) across the three identified stable biological clusters. A = Asthma; C=COPD. P is the p-value for geometric mean difference between cluster 1 or cluster 3 versus cluster 2 (overlap).

5.6 Discussion

In this section, the biological clusters (which were identified in the asthma and COPD test (original) study in chapter 4) were validated using an independent study; and three validation subgroups, which have similar cell-counts and inflammatory mediator patterns as the test study clusters, were identified. The similarity between the cytokine profiles (inflammatory mediators) and cell-counts in test and validation groups supports the view that each cluster is a consistent phenotype and might reflect phenotype-specific responses to treatment.

Two approaches were used to validate the clusters in an independent subgroups using discriminant analysis and the generation of a classifier that used the disease allocation (asthma or COPD status) and sputum IL-1β cutoff. Sputum IL-1β was the best discriminator between the subjects with asthma or COPD in clusters 1 and 3, respectively, with those in the overlap group (cluster 2). Although, validating using linear discriminant analysis is mathematically plausible, the clinical diagnosis of asthma or COPD together with a single sputum cytokine (IL-1β cutoff) demonstrates a simple approach to segment asthma and COPD populations into three groups with distinct and consistent cytokine profiles. This approach, with further validation and study, has advantages in its simplicity and offers the potential for immediate use in stratified medicine studies although it might underestimate small, albeit potentially important subgroups such as T_H2 high COPD. In addition, TNFα shows similar performance as IL-1β in discriminating the overlap group, and could be an alternative potential biomarker for future subjects assignment.

Chapter 6

Modeling Asthma and COPD Biological Heterogeneity at Exacerbation State

6.1 Objectives

The objective of this chapter is to model the biological heterogeneity of asthma and COPD at exacerbation state using appropriate statistical techniques. Subsequently to identify the common and distinctive biological subgroups of both diseases at exacerbation state. The patterns of demographic, clinical, sputum mediators and the microbiome communities will be assessed across the identified biological subgroups/clusters.

6.2 Introduction

In this study, patients with asthma and COPD were followed up at stable state and during exacerbations, with sampling performed in longitudinal visits and in treatment naïve exacerbations. To date this is the largest study that has used biomarker cytokines sampling in longitudinal follow-up and exacerbation visits. An exacerbation is a state in which an asthmatic or COPD patient symptoms get worse and are not enough to control symptoms using the standard treatments such as bronchodilators (inhalers) and steroids. The main symptoms are chest tightness, rapid progressive dyspnea (shortness of breath), dry cough, and extreme wheezing. At worse, it is a life-threatening episode of airway obstruction and is considered as medical emergency state [120].

In this study patient recruitment was performed after fulfilment of specific entry

criteria detailed in chapter 2. An exacerbation was defined according to Anthonisen criteria, and diary cards were used to trigger contact to the research department. Similar to the previous exacerbation studies [121–123], all patients were asked at study entry to contact the research department whether there was an increase in symptoms of breathlessness, sputum production or sputum purulence in comparison to stable state. Following review by a clinician and demonstration that there was no an alternative cause for symptom change (using clinical examination), exacerbation data (such as clinical characteristic and sputum mediators) were captured in those patients who required treatment with systemic corticosteroids and/or antibiotic therapy.

In the previous chapters, the characteristics of asthma and COPD subjects were assessed at stable state, and the common and distinctive biological subgroups were identified. In this chapter, the biological subgroups of asthma and COPD will be investigated at exacerbation state, and they will be assessed whether the patterns are similar or different to the identified stable subgroups. However, first descriptive analysis will be preformed to assess the patterns of the clinical characteristics, sputum mediators (cytokines) and microbiome communities across asthma and COPD at exacerbation state. The patterns might aid to justify that the identified clusters may provide further information which are missing at disease level.

6.3 Study Population

In these prospective studies, thirty-seven asthmatics and seventy-five COPD subjects have exacerbated. At their exacerbation visit, their clinical characteristics such as lung-function (pre- and post-FEV_1), cell-counts (sputum and blood eosinophils, neutrophils, and sputum total cell-count), visual analogue scores (cough, dyspnea), a number of sputum cytokines, and a panel of microbiome communities at both phylum and genus levels were recorded.

6.4 Descriptive Analysis

6.4.1 Patterns of Clinical Characteristics Across Asthma and COPD

The patterns of the clinical parameters were assessed across asthma and COPD at exacerbation state, and presented as summary statistics in table 6.1 on page 109.

Table 6.1: Statistical summaries of demographic and clinical characteristics across asthma and COPD that represent the differences and similarities between the two diseases at exacerbation state.

Variable	Asthma (n=37)	COPD (n=75)	P-value
Pre FEV_1	1.9 (0.15)	1.13 (0.06)	< 0.0001
Post Pre FEV_1	1.99 (0.18)	1.16 (0.06)	< 0.0001
Pre FEV_1 predicted (%)	67.87 (4.63)	43.32 (2.09)	< 0.0001
Post FEV_1 predicted (%)	71.21 (5)	44.5 (2.1)	< 0.0001
Sputum neutrophil count (%)	62.79 (5.03)	74.55 (2.61)	0.03
Blood neutrophils ($x10^9/L$)	6.44 (0.42)	6.85 (0.36)	0.49
VAS score-cough (mm)	63.06 (3.72)	64.18 (2.61)	0.81
VAS score-dypsonea (mm)	65.03 (3.31)	70.16 (2.59)	0.24
Blood eosinophils ($x10^9/L$)$^+$	0.17 (0.12 - 0.24)	0.16 (0.13 - 0.2)	0.92
Sputum eosinophils count (%)$^+$	1.19 (0.56 - 2.52)	1.06 (0.71 - 1.57)	0.77
Macrophage count (%)$^+$	15.14 (10.04 - 22.82)	10.56 (8.04 - 13.86)	0.16
TCC ($x10^6 cells/g\ sputum$)$^+$	3.78 (2.19 - 6.52)	6.28 (4.58 - 8.62)	0.1

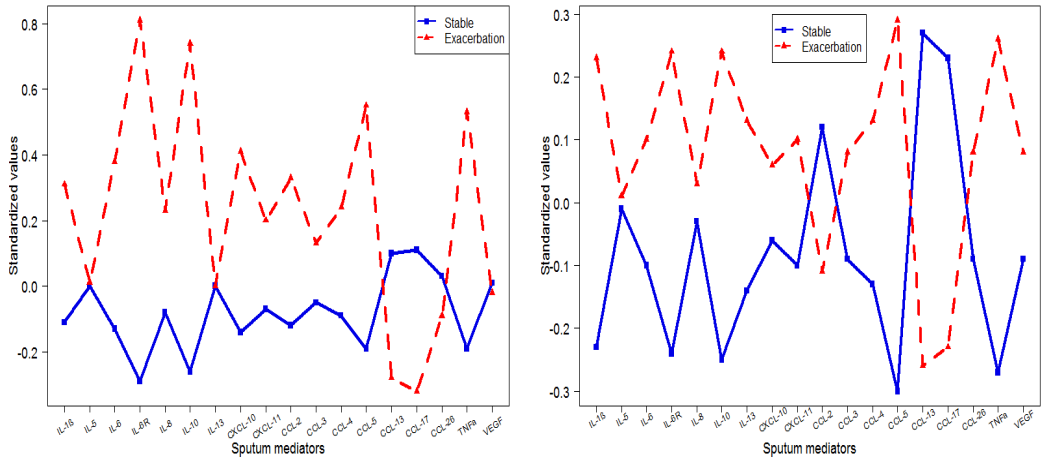
Definition of abbreviations: VAS= Visual Analogue Score; FEV_1 = Forced Expiratory Volume in the First Second; TCC=Total sputum cell count. Data presented as Mean (standard error of mean (SEM)) unless stated; $^+$geometric mean (95% CI)

As shown in the above table, the lung function measurements are significantly lower in COPD subjects compared to asthmatics. However, there is no significant difference between the two diseases in cell-counts and visual analogue scores except in sputum neutrophils.

6.4.2 Patterns of Sputum Mediators Across Asthma and COPD

Thirty one asthmatic and seventy three COPD subjects have records on a number of sputum cytokines (mediators) at exacerbation state. Since, there are corresponding

record for the mediators at stable state, their patterns across stable and exacerbation states were assessed, and presented graphically, separately for asthmatic and COPD subjects in figures 6.1(a) and 6.1(b) on page 110, respectively.



(a) (b)

Figure 6.1: Sputum mediators at stable and exacerbation states: (a) Asthma and (b) COPD

As we observed in the figures above, most of the mediators (in both diseases) appeared to elevate at exacerbation state (except those T_H2 derived mediators) compared to stable state. Sputum IL-6R and CCL-5 were the best discriminators of exacerbation from stable state in asthmatic (figure 6.1(a)), and in COPD subjects (figure 6.1(b), respectively.

In addition, the patterns of these mediators were assessed at exacerbation state across asthma and COPD, and displayed graphically in figure 6.2 on page 111.

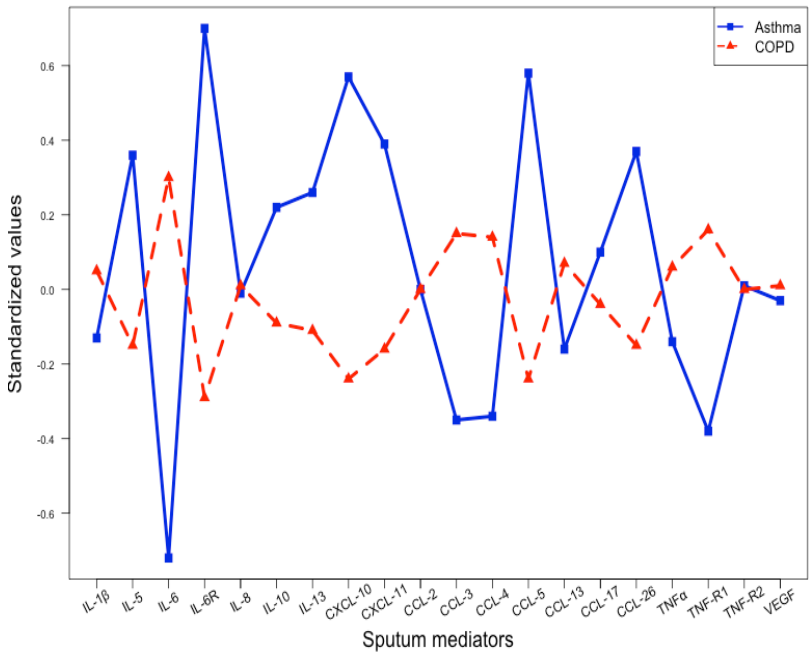


Figure 6.2: Patterns of sputum mediators across asthma and COPD at exacerbation state

As it is illustrated in the figure above, several mediators (such as IL-5, IL-6R, CXCL-10, CXCL-11, CCL-5 and CCL-26) are significantly elevated in asthmatic compared to COPD subjects at exacerbation state. In contrast, IL-6, CCl-3, CCL-4 and TNF-R1 are significantly increased in COPD compared to asthmatic subjects. However, these two diseases do not have significant differences in IL-β, IL-8, IL-10, IL-13, CCL-2, CCL-13, CCL-17, TNFα, TNF-R2 and VEGF, although these mediators show increasing patterns towards specific disease (asthma or COPD) as demonstrated in figure 6.2 on page 111.

Furthermore, to understand the overall patterns of the mediators at exacerbation state across asthma and COPD, principal component analysis (PCA) was performed on the cytokines (displayed in figure 6.2), to reduce to low dimensional components. The first two PCA components (which account for most of the variance of the mediators) were extracted and used to display the data graphically in figure 6.3 on page 112.

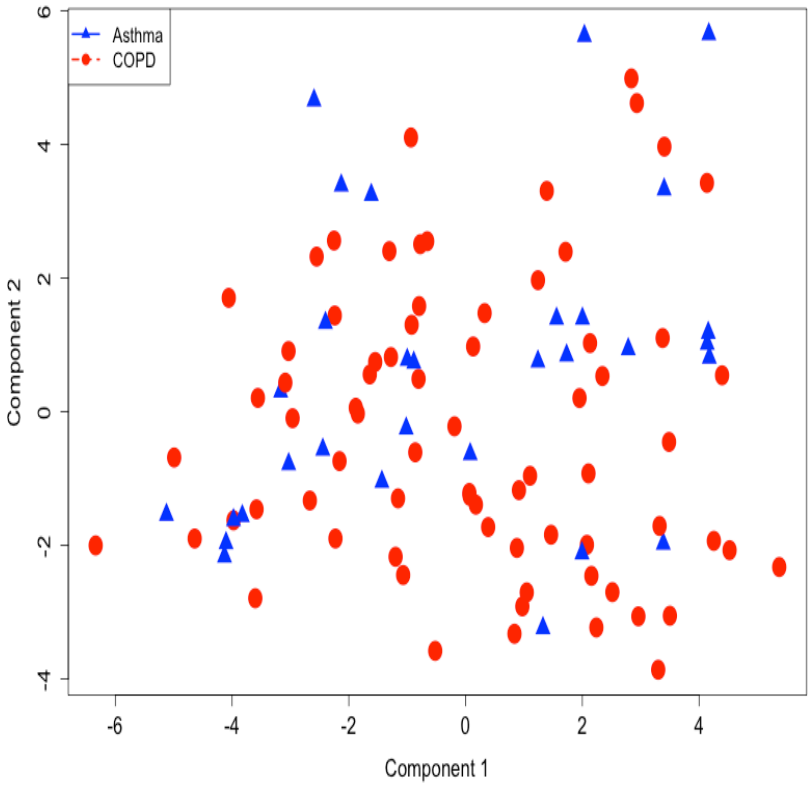
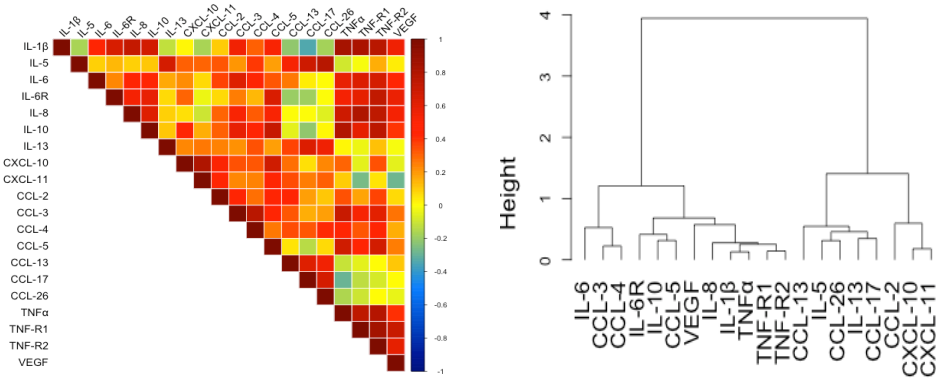


Figure 6.3: Sputum mediators across asthma and COPD at exacerbation state presented using the first two principal component scores

As is shown in the figure above, there is a considerable overlap between asthma and COPD subjects with respect to their sputum mediators at exacerbation state. This observation is similar to the patterns observed at stable state (see figure 2.4 on page 48 in chapter 2).

In addition, to investigate the internal structures (hidden patterns) of the mediators at exacerbation state, the correlations between the mediators are displayed as a heatmap in figure 6.4(a). Further visualization was also performed on these mediators to assess whether they create distinctive subgroups based on their correlation structures, and graphically presented in figure 6.4(b) on page 113.



(a) (b)

Figure 6.4: Sputum mediators at exacerbation state: (a) correlation matrix and (b) subgroups. Heatmap colors: Dark-red indicates strong positive correlation; dark-blue for strong negative correlation; light-red for weak positive correlation; light-blue for weak negative correlation; and yellow for no correlation.

As we observe in the above heatmap there is strong correlations between the mediators, which is similar to the patterns observed at stable state in figure 2.5(a) in chapter 2. In addition, the mediators at exacerbation appeared to create subgroups based on their correlation matrix, and those cytokines which strongly correlated were grouped together. For example, T_H2 derived mediators (IL-5, IL-13, CCL-13, CCL-17, and CCL-26) aggregated together; pro-inflammatory mediators (IL-1β, IL-6, IL-8, IL-10, TNFα, TNF-R1, TNF-R2 and VEGF) grouped together, and T_H1 derived mediators (CXCL-10 and CXCL-11) also formed another group. These patterns are also similar to the patterns observed at stable state in figure 2.5(b) on page 49 in chapter 2.

6.4.3 Patterns of Microbiome Communities Across Asthma and COPD

The pattern of micrbiome profiles were not compared previously across asthma and COPD at exacerbation state. In this study, the 16S rDNA based bacterial community patterns will be assessed across asthma and COPD, and across the identified biological clusters (the differences between the biological subgroups of asthma and COPD might be related to changes in microbial community patterns) at exacerbation visit.

Subjects

In this study, 16 asthmatic and 54 COPD subjects who have sputum mediators recorded at exacerbation state have corresponding microbiome measurements. The microbiome communities were obtained from 16S rRNA sequencing and OTU classification bacterial genomic DNA, which was extracted from the sputum samples using the Qiagen DNA Mini kit (Qiagen, CA, USA) (for details see chapter 2). The sequencing reads were processed using QIIME pipeline [124], and 30 communities (species) at phylum and 399 species at genus levels were identified. In this descriptive analysis, the alpha and beta diversity, and patterns of the most abundant communities will be assessed across asthma and COPD at phylum and genus levels.

6.4.4 Alpha and Beta Diversity of Microbiome Communities

Alpha (α), within a subject, [125] and beta (β), between subjects, [126] diversity at both phylum and genus levels were calculated using Shannon-Weiner and Sorensen indices, respectively, using the Vegan R-package version 2.3 [127].

Alpha diversity is estimated using Shannon-Weaver index, and formulated as follows:

$$H = -\sum_{i=1}^{S} p_i \, log_e(p_i) \tag{6.1}$$

Where, p_i is the proportion of species i, and S is the number of species, so that $\sum_{i=1}^{S} p_i = 1$, e is the base of natural logarithm.

Beta diversity was estimated using Sorensen index, and formulated as follows:

$$\beta = \frac{a+b}{2a+b+c} \tag{6.2}$$

Where, a is the shared species between two subjects, b and c are the unique species in each subject. Then the overall beta diversity is estimated as the mean of all pairwise comparison of the subjects.

Alpha Diversity Across Asthma and COPD

Alpha diversity for each subject was estimated at phylum and genus levels, and the patterns are presented graphically across asthma and COPD subjects in figure 6.5 on page 115.

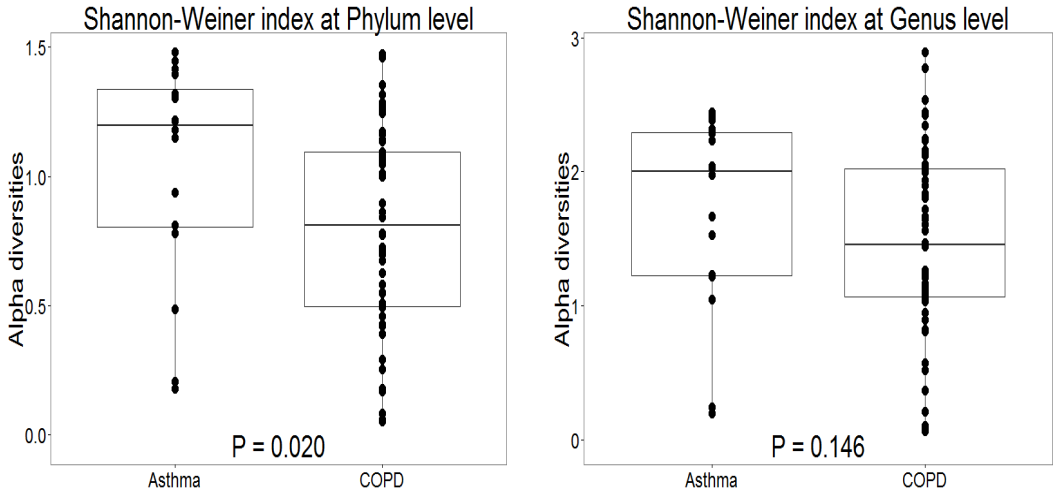


Figure 6.5: Patterns of alpha diversity of microbiome communities at phylum and genus levels across asthma and COPD at exacerbation state

As shown in the figure above, alpha diversity are significantly higher in asthmatic compared to COPD subjects at phylum level. However, there is no significant difference at genus levels between the two diseases.

Beta Diversity Across Asthma and COPD

Beta diversity (as formulated above) was calculated at phylum and genus levels for asthmatic and COPD subjects separately. Beta diversity at phylum level is quite similar between the two diseases, in which 0.20 in asthmatic and 0.21 in COPD subjects. However, it is higher in COPD at genus level, in which 0.43 in asthmatic and 0.51 in COPD subjects.

6.4.5 Patterns of the Most Abundant Microbiome Communities Across Asthma and COPD

The relative abundance of each community was calculated at phylum and genus levels, and if any of the species has more than 2% median relative abundance in either asthma and/or COPD diseases were used for further analysis. Therefore, Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria at phylum level; and Actinomyces (phylum Actinobacteria), Rothia (phylum Actinobacteria), Prevotella (phylum Bacteroidetes), Gemella (phylum Firmicutes), Streptococcus (phylum Firmicutes), Veillonella (phylum Firmicutes) and Haemophilus (phylum Proteobacteria) at genus level satisfied the criteria. The patterns of these highly abundant species were investigated across asthma and COPD, and graphically presented.

Patterns at Phylum Level Across Asthma and COPD

The patterns of the four most abundant microbiome species at phylum level (Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria) are presented graphically across asthma and COPD in figure 6.6 on 117.

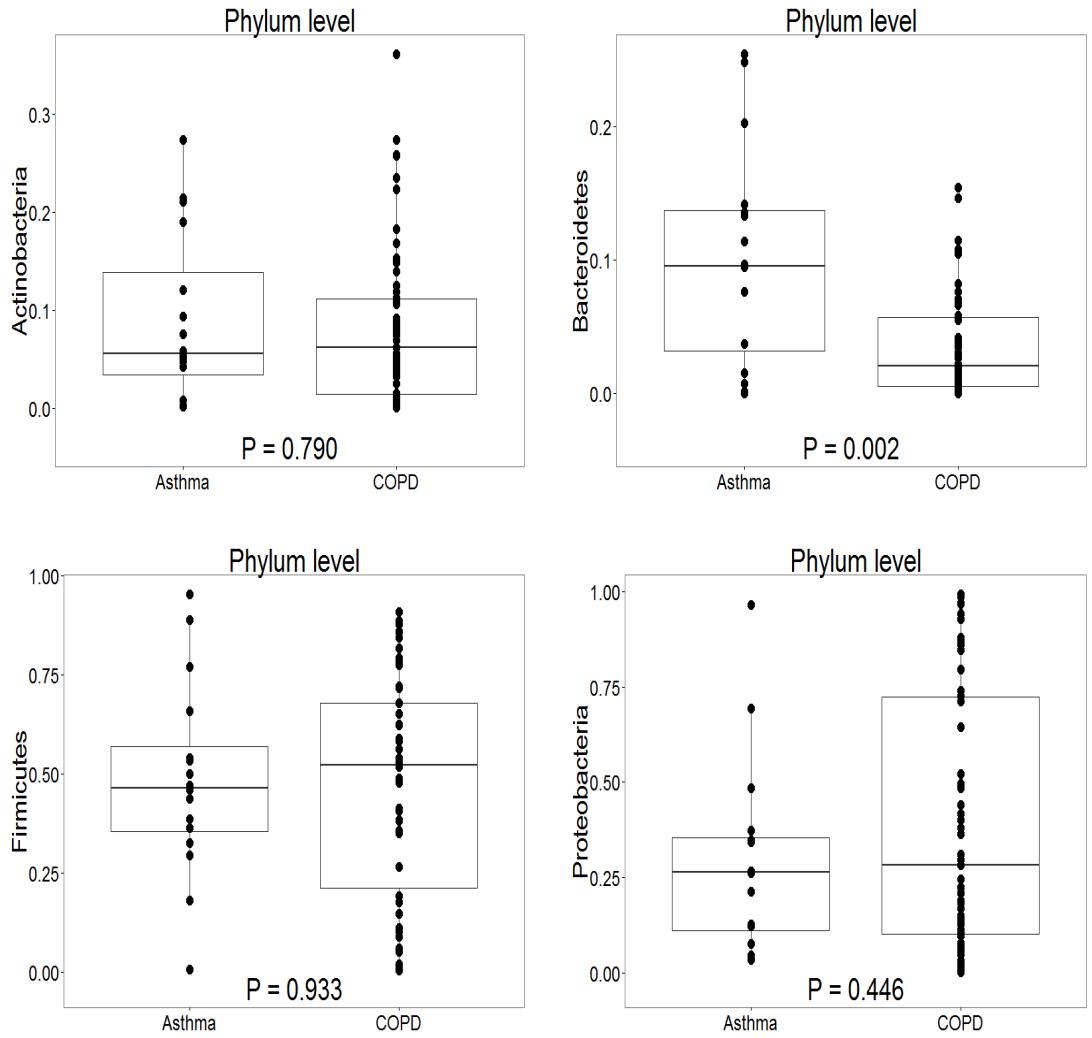


Figure 6.6: Patterns of microbiome communities at phylum level across asthma and COPD at exacerbation

As observed in the figure above, only Bacteroidetes are significantly elevated in asthmatics compared to COPD subjects. However, there is no clear difference in Actinobacteria, Firmicutes and Proteobacteria across the two diseases.

Patterns at Genus Level Across Asthma and COPD

The patterns of the most abundant genera such as Actinomyces (phylum Actinobacteria), Rothia (phylum Actinobacteria), Prevotella (phylum Bacteroidetes), Gemella (phylum Firmicutes), Streptococcus (phylum Firmicutes), Veillonella (phylum Firmicutes) and Haemophilus (phylum Proteobacteria) are graphically presented across asthma and COPD subjects in figure 6.7 on page 118.

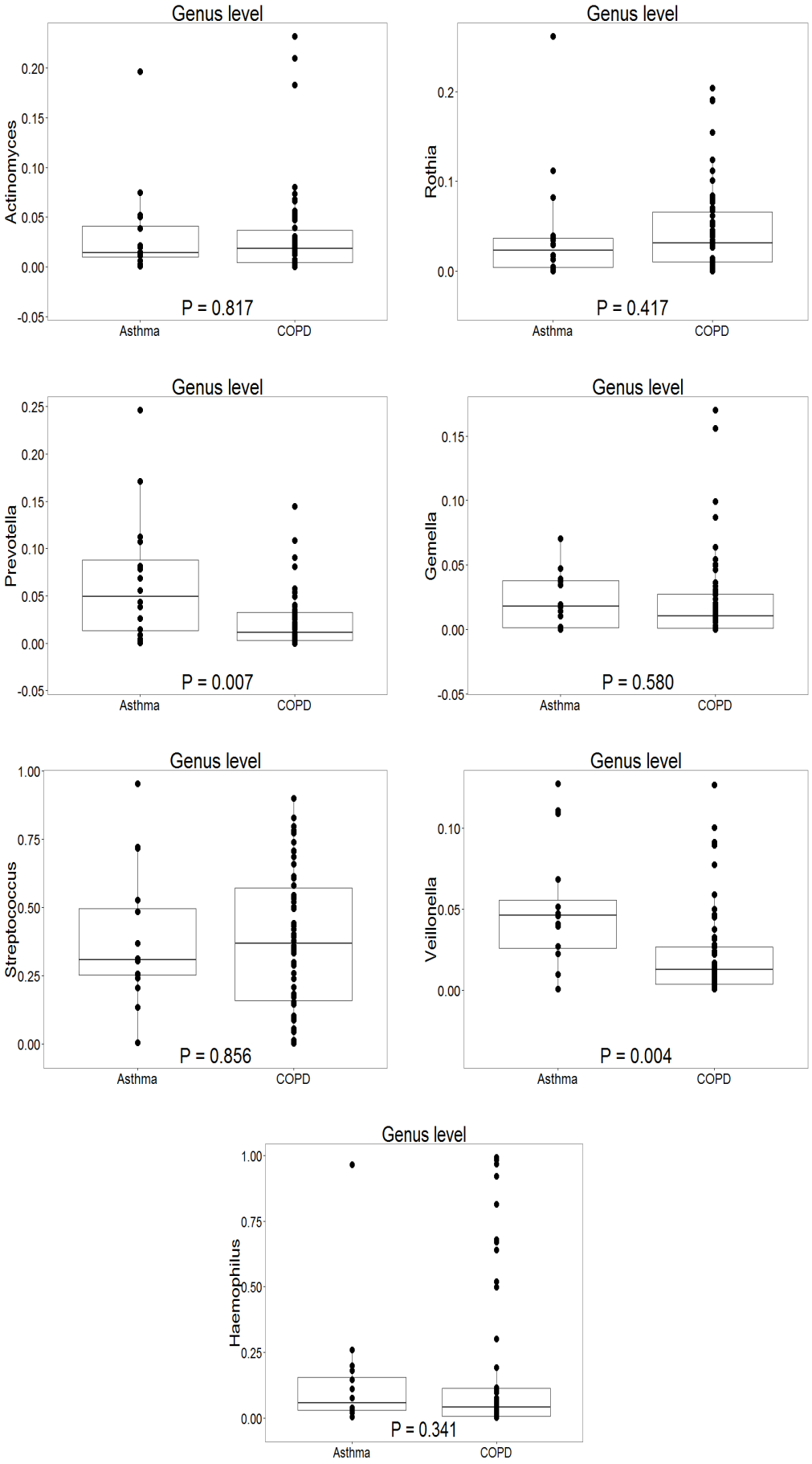


Figure 6.7: Pattern of microbiome communities at genus level across asthma and COPD at exacerbation

As it has been illustrated in the figure above, only Prevotella (phylum Bacteroidetes) and Veillonella (phylum Firmicutes) are significantly elevated in asthmatic compared to COPD subjects at exacerbation state. However, there are no significant differences across the diseases in the others most abundant genera (such as Actinomyces, Rothia, Gemella, Streptococcus and Haemophilus) although they show sign of elevation towards specific disease.

6.4.6 Descriptive Summaries

In the above descriptive analysis, we observed that several clinical variables, sputum mediators and most abundant communities did not show any significant difference across asthma and COPD at disease level. Therefore, these patterns will be assessed further across the exacerbation biological subgroups, in order to reveal any hidden patterns which are missing when only comparing at disease level.

6.5 Statistical Methods for Biological Clustering

From the above explanatory analysis, we can observe that the patterns of the sputum mediators at exacerbation are very similar to the patterns observed at stable state. Therefore, a two stage (factor and cluster analyses) was performed to identify the common and distinctive exacerbation biological clusters/subgroups of asthma and COPD.

First, factor analysis (with varimax rotation) was applied to all the cytokines and reduced to small independent factors. The optimal factors were retained on the basis of screeplot (factors above the break in the curve) and eigenvalue above one [102]. Subsequently, the corresponding factor scores which represent each subject was generated and used as input into k-means cluster analysis to identify the subjects clusters (for the statistical methods details readers refer to chapter 3 and 4). The statistical summary of all the available demographic and clinical characteristics were presented across the biological subgroups. In addition, alpha and beta diversity, and the patterns of the most abundant communities (at both phylum and genus levels) are presented across the biological clusters. The biological subgroups (clusters) were interpreted according to these characteristics patterns.

In addition, after the biological clusters were identified using factor scores as input to k-means clustering, linear discriminant analysis was applied to predict the subgroups using the actual cytokines measurements. Then mediators which have substantial contribution in discriminating the subgroups were identified, and number of clusters less one discriminant function scores for each subject were generated and used to display the subjects subgroups graphically.

6.6 Results

6.6.1 Asthma and COPD Factors at Exacerbation State

Factor analysis (principal factor with varimax rotation) was performed on asthma and COPD mediators at exacerbation state, and four factors were retained which accounted for about 94.2% of the total shared variance and almost all the correlation that exists between the mediators. Since IL-13 has below limit of detection in most of the subjects, it was excluded from factor analysis; however, its pattern was assessed across the identified subgroups in the post analysis. The rotated factor loadings are depicted in table 6.2 on page 121.

Table 6.2: Varimax rotated factor loadings of sputum mediators at exacerbation

	Factor 1	Factor 2	Factor 3	Factor 4	C^1	U^2
IL-1β	0.90	-0.25	-0.06	0.10	0.88	0.12
IL-5	0.04	0.87	0.26	0.02	0.83	0.17
IL-6	0.50	0.01	0.07	0.57	0.57	0.43
IL-6R	0.82	0.13	0.19	-0.30	0.82	0.18
IL-8	0.80	0.02	-0.03	0.17	0.67	0.33
IL-10	0.68	-0.08	0.38	0.15	0.64	0.36
CXCL-10	0.07	0.23	0.90	0.08	0.88	0.12
CXCL-11	-0.11	0.24	0.83	0.15	0.77	0.23
CCL-2	0.11	0.23	0.50	0.40	0.47	0.53
CCL-3	0.50	0.16	0.27	0.66	0.78	0.22
CCL-4	0.39	0.34	0.18	0.57	0.62	0.38
CCL-5	0.63	0.03	0.64	-0.03	0.82	0.18
CCL-13	-0.12	0.55	0.17	0.43	0.54	0.46
CCL-17	-0.24	0.81	0.02	0.23	0.76	0.24
CCL-26	-0.06	0.84	0.20	-0.04	0.75	0.25
TNFα	0.82	-0.27	0.22	0.30	0.89	0.11
TNF-R1	0.91	-0.09	-0.20	0.19	0.92	0.08
TNF-R2	0.92	0.05	0.17	0.16	0.91	0.09
VEGF	0.67	0.09	-0.26	0.07	0.52	0.48
Eigenvalue	6.54	2.91	2.77	1.83		

C^1 = Proportion of total variation accounted for by the common factors (common variance)
U^2 = Proportion of total variation not accounted by the common factors (unique variance)

In the table above, clear patterns in the factor loadings were observed, in which proinflammatory mediators appeared to load in factor 1, T_H2 mediators in factor 2, and T_H1 mediators in factor 3. These observations are quite consistent with the patterns observed in figure 6.4 (b) on page 113.

6.6.2 Asthma and COPD Clusters at Exacerbation State

In this study, three biological clusters were identified at exacerbation state using factor scores as input into k-means clustering. The clinical parameters and mediators are presented across the identified subgroups in table 6.3 and table 6.4, respectively. In addition, the clusters are presented graphically across the first two discriminant functions scores (number of clusters less one) in figure 6.8 on page 123; and the patterns of the mediators across the subgroups are also presented in figure 6.9 on page 123.

Table 6.3: Statistical summaries of demographic and clinical characteristics across the three identified biological clusters at exacerbation that represent the differences and similarities between the clusters

Variable	Cluster 1	Cluster 2	Cluster 3	P-value C1 vs. C2	P-value C1 vs. C3	P-value C2 vs. C3
Male [n (%)]	21 (65.6)	21 (63.6)	24 (61.5)	0.87	0.72	0.85
Current or Ex-smokeres [n (%)]	25 (78.1)	21 (63.6)	31 (79.5)	0.2	0.9	0.13
Age (year)	63 (2.0)	63 (2.4)	67 (1.6)	0.34	0.95	0.36
BMI (kg/m^2)	28.4 (1.2)	28.9 (1.1)	25.5 (0.72)	0.47	0.96	0.48
Pre FEV_1 (L)	1.37 (0.15)	1.52 (0.12)	1.14 (0.09)	0.39	0.18	0.01
Post FEV_1 (L)	1.43 (0.18)	1.48 (0.13)	1.22 (0.09)	0.79	0.25	0.09
Pre FEV_1 predicted (%)	47.43 (4.1)	54.7 (3.87)	46.38 (3.81)	0.23	0.85	0.14
Post FEV_1 predicted (%)	49.01 (4.86)	54.33 (4.06)	49.05 (3.7)	0.4	0.99	0.35
Sputum neutrophil count(%)	60.44 (3.82)	59.89 (4.54)	87.36 (2.52)	0.9	<0.0001	<0.0001
Blood neutrophils (x$10^9/L$)	5.88 (0.42)	6.26 (0.48)	8.02 (0.54)	0.59	0.002	0.02
VAS score-cough (mm)	65 (4.11)	63.24 (3.77)	66.67 (3.21)	0.74	0.75	0.48
VAS score-dypsonea (mm)	72.19 (4.12)	62.88 (3.87)	68.77 (3.11)	0.08	0.51	0.23
Blood eosinophils (x$10^9/L$)	0.29 (0.21 - 0.4)	0.13 (0.09 - 0.17)	0.13 (0.1 - 0.17)	<0.0001	<0.0001	0.83
Sputum eosinophils count (%)	5.6 (2.95 - 10.64)	0.75 (0.43 - 1.3)	0.39 (0.3 - 0.5)	<0.0001	<0.0001	0.02
Macrophage count (%)	16.5 (12.4 - 21.94)	22.61 (15.98 - 32.0)	5.71 (3.95 - 8.25)	0.22	<0.0001	<0.0001
TCC (x$10^6 cells/g\, sputum$)	3.09 (2.1 - 4.55)	3.38 (2.01 - 5.66)	12.44 (8.24 - 18.8)	0.79	<0.0001	<0.0001
Bacterial colonization [n/N (%)]	7/30 (23.3)	8/26 (30.8)	26/38 (68.4)	0.53	<0.0001	0.003

Definition of abbreviations: VAS= Visual Analogue Score; BMI= Body Mass Index; FEV_1 = Forced Expiratory Volume in the First Second; TCC=Total sputum cell count, C=cluster; Cluster 1= (Asthma=11 and COPD=21); Cluster 2= (Asthma=15 and COPD=18); Cluster 3 = (Asthma=5 and COPD=34). Data presented as geometric mean (95% CI) unless stated;$^+$Mean (standard error of mean (SEM)); *median (1^{st} and 3^{rd} quartiles); Dose for only those subjects prescribed daily prednisolone; Pack-year history of current and ex-smokers; ^{a}beclomethasonedipropionate equivalent

Table 6.4: Statistical summaries of sputum mediators across the three identified biological clusters at exacerbation state that represent the differences and similarities between the clusters

Variable	Cluster 1	Cluster 2	Cluster 3	P-value C1 vs. C2	P-value C1 vs. C3	P-value C2 vs. C3
IL-1β	42.2 (23.3 - 76.4)	92.6 (49.4 - 173.5)	1093 (616.1 - 1938.8)	0.08	<0.0001	<0.0001
IL-5	5.8 (3.7 - 9.2)	1.7 (1 - 2.9)	0.6 (0.5 - 0.8)	<0.0001	<0.0001	<0.0001
IL-6	160.5 (95.8 - 269)	368.5 (186.1 - 729.7)	708 (435.2 - 1151.8)	0.05	<0.0001	0.12
IL-6R	270.1 (161 - 452.9)	305.6 (192.2 - 485.9)	640.4 (424.1 - 967.1)	0.72	0.01	0.02
IL-8	2929 (2045 - 4195.3)	3121.6 (2012.7 - 4841.4)	10493 (7938.8 - 13868.9)	0.81	<0.0001	<0.0001
IL-10	2.1 (1.8 - 2.6)	12.9 (6.4 - 25.9)	10.8 (6.6 - 17.7)	<0.0001	<0.0001	0.68
IL-13	14.1 (11.2 - 17.6)	11.1 (9.3 - 13.3)	8.9 (8 - 9.9)	0.07	<0.0001	0.03
CXCL-10	383.9 (236.4 - 623.3)	3740.1 (2271.6 - 6158.1)	147.4 (93.5 - 232.2)	<0.0001	0.006	<0.0001
CXCL-11	24.5 (13.7 - 43.9)	442.1 (174.2 - 1121.9)	4.4 (2.8 - 6.8)	<0.0001	<0.0001	<0.0001
CCL-2	381.6 (268 - 543.4)	947.2 (581.8 - 1541.8)	297.7 (225.5 - 392.9)	0.001	0.36	<0.0001
CCL-3	44.7 (29.7 - 67.2)	72.6 (40.3 - 130.8)	88.8 (59.7 - 132)	0.17	0.04	0.57
CCL-4	976.4 (625.1 - 1525.1)	1072.1 (573.4 - 2004.5)	1114.5 (750.2 - 1655.8)	0.8	0.71	0.92
CCL-5	5 (3.3 - 7.7)	19.7 (12.1 - 32.1)	9.9 (6.9 - 14.3)	<0.0001	0.03	0.03
CCL-13	27.8 (22 - 35)	20 (13.9 - 28.8)	12.1 (10.1 - 14.4)	0.1	<0.0001	0.01
CCL-17	50.9 (32.8 - 79)	8.6 (5.7 - 13)	4.7 (3.1 - 7.2)	<0.0001	<0.0001	0.05
CCL-26	18.6 (12.6 - 27.5)	5.2 (3.4 - 7.7)	2.3 (1.8 - 2.8)	<0.0001	<0.0001	<0.0001
TNF-α	2.5 (1.5 - 4.1)	20.9 (9.5 - 45.8)	78.3 (45.6 - 134.4)	<0.0001	<0.0001	0.007
TNF-R1	754.8 (562.4 - 1013.1)	772.6 (528.5 - 1129.4)	4143.4 (3157.4 - 5437.3)	0.92	<0.0001	<0.0001
TNF-R2	303.7 (206.9 - 445.9)	564.9 (328.3 - 972.2)	1387.9 (980 - 1965.6)	0.05	<0.0001	0.006
VEGF	1212.7 (1021.5 - 1439.8)	1083.2 (904.8 - 1296.9)	2035.7 (1652.4 - 2508.1)	0.43	<0.0001	<0.0001

Definition abbreviations: C=cluster; Cluster 1= (Asthma=11 and COPD=21); Cluster 2= (Asthma=15 and COPD=18); Cluster 3 = (Asthma=5 and COPD=34). Data presented as geometric mean with corresponding 95% confidence interval (CI); unit of the mediators is pg/ml.

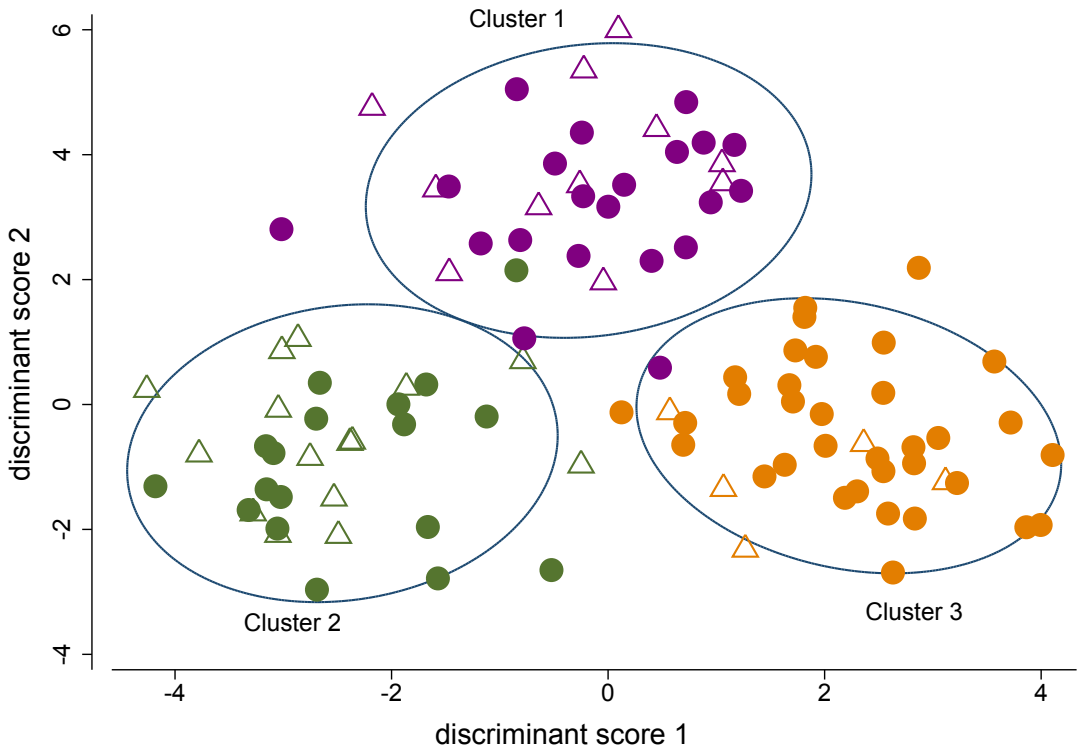


Figure 6.8: The 3 identified exacerbation biological clusters presented using subjects' discriminant scores. Hollow triangle indicates asthma and bold circle indicates COPD.

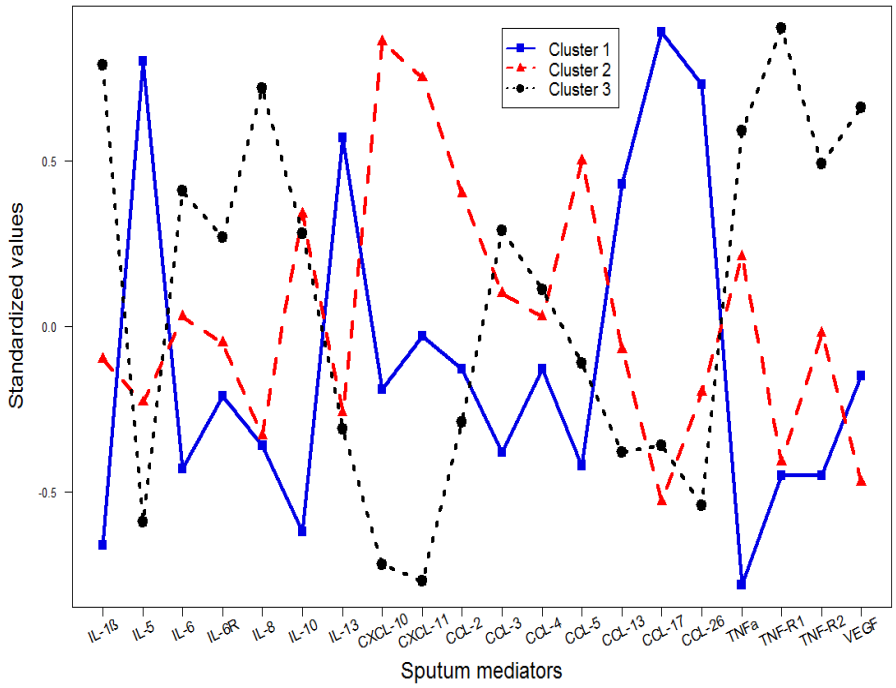


Figure 6.9: Patterns of sputum mediators across the three clusters of asthma and COPD at exacerbation

Cluster 1

Cluster 1 consists of 31 subjects (asthma =11 and COPD =21), in which 65% are men, with 65 years average age, and 78% are current or ex-smokers. Subjects in this cluster have high blood and sputum eosinophil count, 5.6 (95%: 2.95 - 10.64), and 0.29 (95%: 0.21 - 0.4), respectively (Table 6.3). In addition, these subjects have elevated T_H2 cytokines (IL-5, IL-13, CCL-13, CCL-17 and CCL-26), table 6.4 and figure 6.9.

Subjects in this cluster have also clinical records at stable state, and pairwise comparison were performed to compared the patterns of these characteristics at stable state with their corresponding levels at the exacerbation (table 6.5). The lung function measurements (such as pre- and post-FEV_1, pre- and post FEV_1 percentage predicted) were significantly lower at exacerbation compared to stable state (all p-value < 0.01). In contrast, visual analogue scores (cough and dyspnea) (all p-value < 0.0001), and sputum eosinophils (p-value $= 0.02$) were significantly higher at exacerbation compared to stable state.

Table 6.5: Statistical summaries of the pairwise comparison (within subject) of the clinical parameters between stable and exacerbation states in cluster 1

Variable	Stable	Exacerbation	P-value
Pre FEV_1 (L)	1.59 (0.14)	1.37 (0.15)	0.002
Post FEV_1 (L)	1.68 (0.16)	1.43 (0.18)	<0.001
Pre FEV_1 predicted (%)	56.73 (4.15)	47.43 (4.1)	0.004
Post FEV_1 predicted (%)	59.46 (4.43)	49.01 (4.86)	<0.001
Sputum neutrophil count (%)	62.45 (3.87)	60.44 (3.82)	0.45
Blood neutrophils (x$10^9/L$)	5.2 (0.3)	5.88 (0.42)	0.06
VAS score-cough (mm)	32.75 (5.59)	65 (4.11)	<0.0001
VAS score-dypsonea (mm)	35.91 (5.6)	72.19 (4.12)	<0.0001
Blood eosinophils (x$10^9/L$)$^+$	0.27 (0.21 - 0.34)	0.29 (0.21 - 0.4)	0.64
Sputum eosinophils count (%)$^+$	2.39 (1.3 - 4.4)	5.6 (2.95 - 10.64)	0.02
Macrophage count (%)$^+$	22.55 (17.82 - 28.54)	16.5 (12.4 - 21.94)	0.06
TCC (x$10^6 cells/g\ sputum$)$^+$	2.21 (1.55 - 3.16)	3.09 (2.1 - 4.55)	0.23

Definition of abbreviations: VAS= Visual Analogue Score; FEV_1 = Forced Expiratory Volume in the First Second; TCC=Total sputum cell count. Data presented as Mean (standard error of mean (SEM)) unless stated; $^+$geometric mean (95% CI)

In addition, the patterns of the sputum mediators across stable and exacerbation were assessed in this cluster, and graphically illustrated in figure 6.10 on page 125. However, most of the mediators have no significant difference between stable and exacerbation states (except CCL-26) in this cluster.

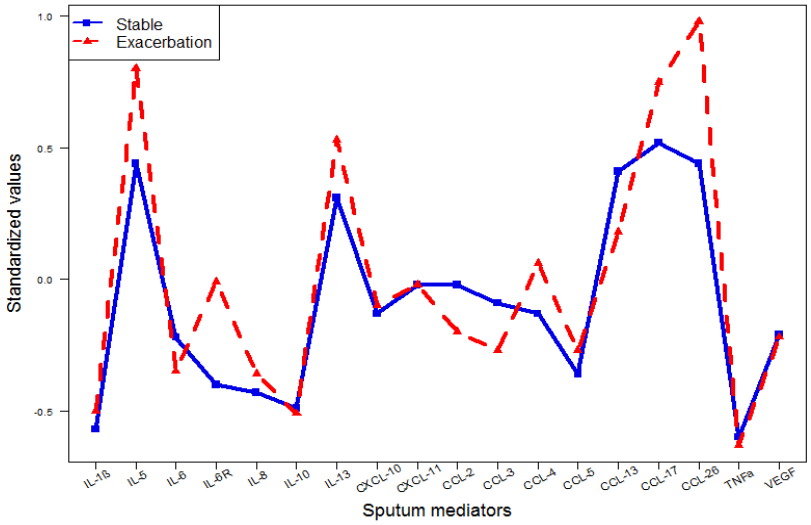


Figure 6.10: Patterns of sputum mediators across stable and exacerbation in cluster 1

Cluster 2

Cluster 2- consists of 33 subjects (asthma=15, COPD=18), 64% are men with 66 years average age, and 64% are current or ex-smokers. The subjects in this cluster have elevated level in T_H1 derived cytokines (CXCL-10, CXCL-11) and CCL-5 compared to cluster 1 (tables 6.3 and 6.4 and figure 6.9).

In addition, the patterns of the clinical characteristics were compared (using pairwise comparison) between stable and exacerbation states in this cluster, and presented in table 6.6. The lung function measurements (such as pre- and post-FEV_1, pre- and post-FEV_1 percentage predicted) were significantly lower at exacerbation compared to stable state (all p-value < 0.01). In contrast, visual analogue scores (cough and dyspnea) were significantly incremented at exacerbation compared to stable state (all p-value <0.0001), which is similar to the pattern observed in cluster 1. However, no blood or sputum cell-counts were significantly different between exacerbation and stable states, which is in contrast to the patterns observed in cluster 1 (see table 6.6).

Table 6.6: Statistical summaries of the pairwise comparison (within subject) of the clinical parameters between stable and exacerbation states in cluster 2

Variable	Stable	Exacerbation	P-value
Pre FEV_1 (L)	1.72 (0.12)	1.52 (0.12)	0.005
Post FEV_1 (L)	1.8 (0.13)	1.48 (0.13)	<0.0001
Pre FEV_1 predicted (%)	62 (4.03)	54.7 (3.87)	0.004
Post FEV_1 predicted (%)	63.84 (3.86)	54.33 (4.06)	<0.001
Sputum neutrophil count (%)	67.94 (3.61)	59.89 (4.54)	0.29
Blood neutrophils ($x10^9/L$)	5.75 (0.37)	6.26 (0.48)	0.21
VAS score-cough (mm)	30.67 (4.28)	63.24 (3.77)	<0.0001
VAS score-dypsonea (mm)	39.61 (4.08)	62.88 (3.87)	<0.0001
Blood eosinophils ($x10^9/L$)$^+$	0.17 (0.12 - 0.26)	0.13 (0.09 - 0.17)	0.2
Sputum eosinophils count (%)$^+$	1.09 (0.68 - 1.75)	0.75 (0.43 - 1.3)	0.12
Macrophage count (%)$^+$	17.28 (11.51 - 25.96)	22.61 (15.98 - 32.01)	0.46
TCC ($x10^6 cells/g\ sputum$)$^+$	1.78 (1.14 - 2.78)	3.38 (2.01 - 5.66)	0.06

Definition of abbreviations: VAS= Visual Analogue Score; FEV_1 = Forced Expiratory Volume in the First Second; TCC=Total sputum cell count. Data presented as Mean (standard error of mean (SEM)) unless stated; $^+$geometric mean (95% CI)

In addition, the patterns of the mediators were assessed (using pairwise comparison) across stable and exacerbation states in this group, and most of the mediators (such as IL-1β, IL-6, IL-10, IL-13, CXCL-10, CXCL-11, CCL-3, CCL-5 and TNFα) appeared to elevated significantly at exacerbation compared to stable state. However, CCL-13 and CCL-17 decreased significantly at exacerbation compared to the stable state (Figure 6.11).

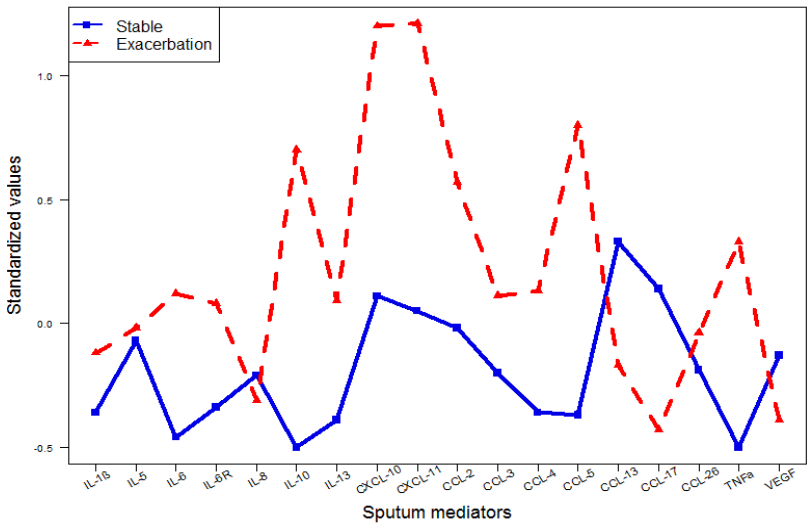


Figure 6.11: Patterns of sputum mediators across stable and exacerbation states in cluster 2

Cluster 3

Cluster 3 is a COPD predominant group which consists of 39 subjects (asthma=5 and COPD=34), 61% are men with 67 average age and 79% were current or ex-smokers. Subjects in this group have elevated level of blood and sputum neutrophil cell counts and high level of pro-inflammatory mediators (IL-1β, IL-6, IL-6R, TNFα, TNF-R1, TNF-R2 and VEGF) and increased bacterial colonisation compared to cluster 1 and 2.

Subjects in this cluster have also clinical records at stable state, and pairwise comparison were performed to compare with their corresponding levels at exacerbation state (table 6.6). The lung function measurements (such as pre- and post-FEV_1, pre- and post FEV_1 percentage predicted) were significantly lower at exacerbation (all p-value < 0.01). Whereas visual analogue scores (cough and dyspnea), and blood and sputum neutrophil and sputum total-cell-count are significantly increased at exacerbation (all p-value <0.05). However, blood or sputum eosinophils cell-counts are significantly lower at exacerbation compared to stable state (all p-value < 0.05), which is in contrast to the patterns observed in cluster 1 (table 6.7).

Table 6.7: Statistical summaries of the pairwise comparison (within subject) of the clinical parameters between stable and exacerbation states in cluster 3

Variable	Stable	Exacerbation	P-value
Pre FEV_1 (L)	1.27 (0.09)	1.14 (0.09)	0.002
Post FEV_1 (L)	1.34 (0.1)	1.22 (0.09)	0.001
Pre FEV_1 predicted (%)	52.26 (4)	46.38 (3.81)	<0.001
Post FEV_1 predicted (%)	54.55 (4.02)	49.05 (3.7)	<0.0001
Sputum neutrophil count (%)	73.25 (3.63)	87.36 (2.52)	0.003
Blood neutrophils (x$10^9/L$)	6.18 (0.43)	8.02 (0.54)	<0.001
VAS score-cough (mm)	49.46 (4.51)	66.67 (3.21)	<0.001
VAS score-dypsonea (mm)	51.61 (3.89)	68.77 (3.11)	<0.001
Blood eosinophils (x$10^9/L$)$^+$	0.19 (0.14 - 0.25)	0.13 (0.1 - 0.17)	0.04
Sputum eosinophils count (%)$^+$	0.96 (0.6 - 1.54)	0.39 (0.3 - 0.5)	<0.001
Macrophage count (%)$^+$	12.63 (9.25 - 17.26)	5.71 (3.95 - 8.25)	0.001
TCC (x$10^6 cells/g\ sputum$)$^+$	4.45 (2.97 - 6.67)	12.44 (8.24 - 18.8)	<0.001

Definition of abbreviations: VAS= Visual Analogue Score; FEV_1 = Forced Expiratory Volume in the First Second; TCC=Total sputum cell count. Data presented as Mean (standard error of mean (SEM)) unless stated; $^+$geometric mean (95% CI)

In addition, the pattern of the mediators in this cluster were assessed across stable and exacerbation using pairwise comparison approach, and displayed in figure 6.12 on page 128. In this cluster, the overall pattern of the mediators across stable and exacerbation states is not quite clear, which is dissimilar to the patterns observed in clusters 1 and 2.

Several mediators (such as IL-1β, IL-6R, IL-8, IL-10, CXCL-10, CXCL-11, CCL-5, TNFα, VEFG) were significantly higher at exacerbation; in contrast, most of the T_H1 and T_H2 mediators (such as IL-5, CXCL-10, CXCL-11, CCL-2, CCL-13, CCL-17 and CCL-26) were significantly lower at exacerbation compared to their levels at stable state.

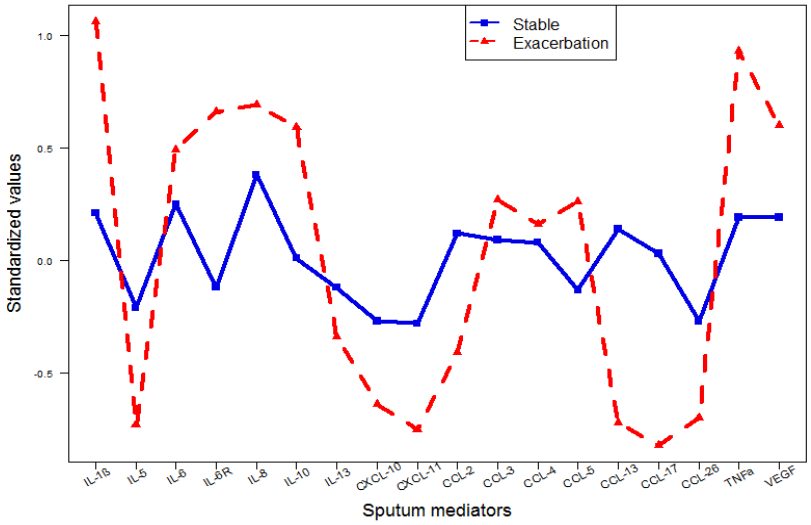


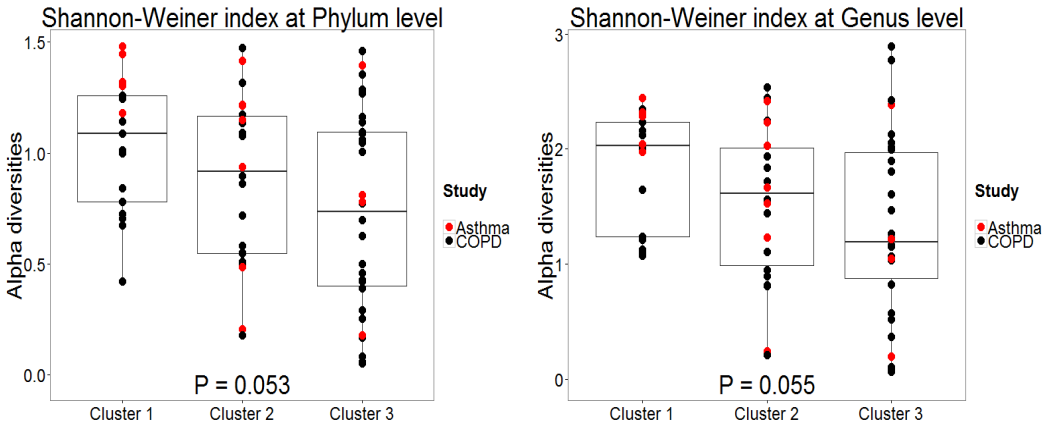
Figure 6.12: Pattern of sputum mediators across stable and exacerbation in cluster 3

6.6.3 Patterns of Microbiome Communities Across the Biological Clusters

Some of the subjects who were assigned to the biological clusters have corresponding microbiome information at both phylum and genus levels; in which 17 subjects (asthma=5; COPD=12) in cluster 1, 22 subjects (asthma=7; COPD=15) in cluster 2, and 30 subjects (asthma=4; COPD=26) in cluster 3 had microbiome records. In this section, the pattern of alpha and beta diversity indices and the most abundant communities will be assessed across the subgroups.

Patterns of Alpha and Beta Diversity Across the Biological Clusters

Alpha diversity at phylum and genus levels are estimated across the biological clusters, and displayed in figure 6.13 on page 129.



(a) (b)

Figure 6.13: Patterns of alpha diversity across the biological clusters at exacerbation state: (a) at phylum level and (b) at genus level

As observed in the figure above, the pattern of alpha diversity is not significantly different between the three biological clusters (although a trend is seen that is higher in cluster 1, then cluster 2 compared to cluster 3).

In addition, beta diversity is estimated for each cluster, which is quite similar across the clusters at phylum level, in which 0.21 in cluster 1, 0.20 in cluster 2 and 0.21 in cluster 3. However, at genus level it appears higher in cluster 3 compared to the other clusters, in which 0.44 in cluster 1, 0.50 in cluster 2, and 0.54 in cluster 3.

Patterns of the Most Abundant Microbiome Communities Across the Biological Clusters at Phylum Level

The patterns of the most abundant microbiome communities at phylum level (Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria) are presented across the biological exacerbation clusters in figure 6.14 on page 130.

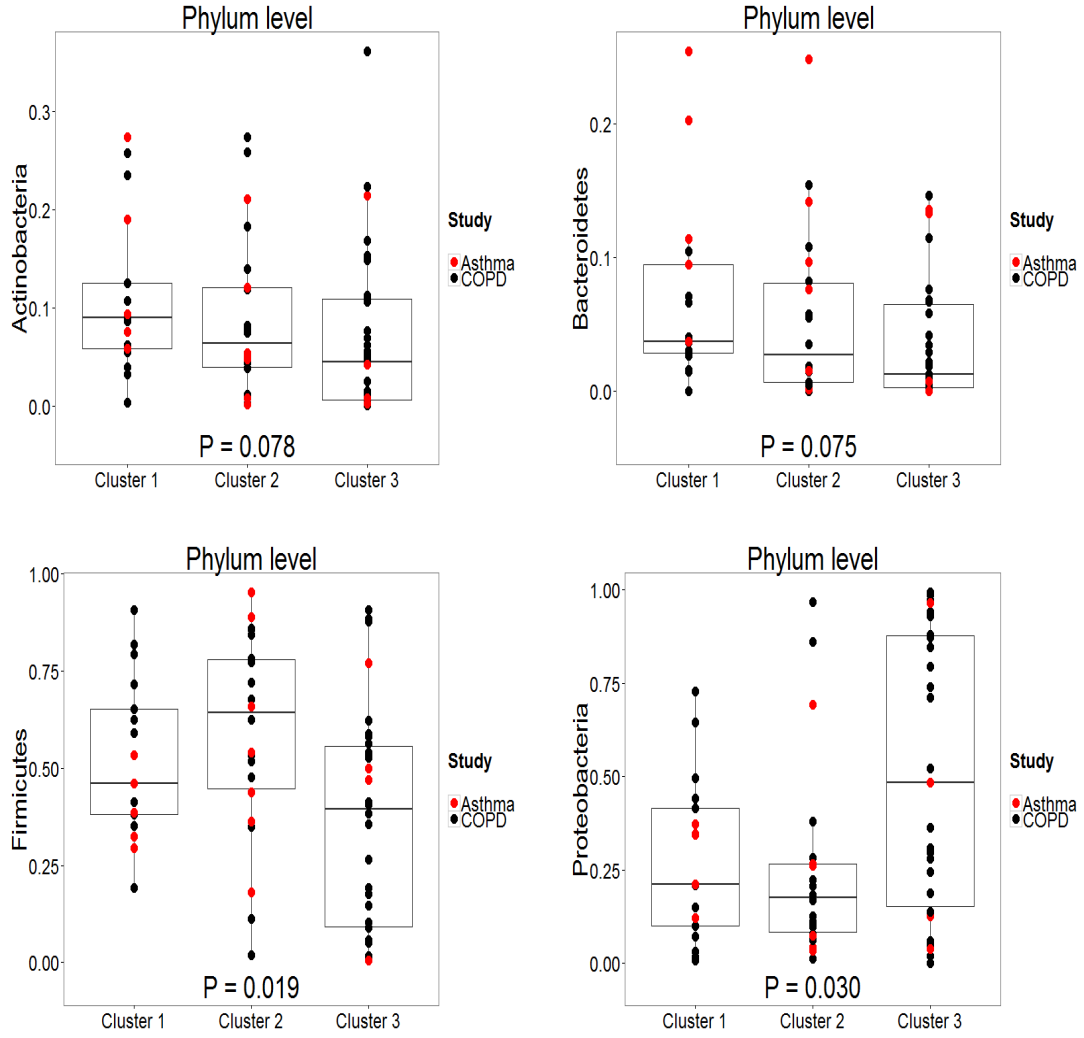


Figure 6.14: Patterns of the most abundant microbiome communities at phylum level across the biological clusters at exacerbation state

As illustrated in the above figure, Firmicutes and Proteobacteria are significantly elevated in cluster 2 and cluster 3, respectively. However, the abundance of Actinobacteria and Bacteroidetes look high in cluster 1 compared to the other clusters, but not statistically significant. These observations are quite novel as the we observed in figure 6.6 on 117, Firmicutes and Proteobacteria are not statistically significant across asthma and COPD (at disease level) but they are significantly different across the biological clusters.

Patterns of the Most Abundant Microbiome Communities Across the Biological Clusters at Genus Level

The patterns of the microbiome communities at genus level such as Actinomyces (phylum Actinobacteria), Rothia (phylum Actinobacteria), Prevotella (phylum Bacteroidetes), Gemella (phylum Firmicutes), Streptococcus (phylum Firmicutes), Veillonella (phylum Firmicutes) and Haemophilus (phylum Proteobacteria) were investigated across the identified exacerbation biological clusters, and are graphically presented in figure 6.15 on page 132.

As demonstrated in the figures below, only Streptococcus (phylum Firmicutes) is significantly elevated in cluster 2 (this is consistent with Firmicutes's pattern at phylum level across the clusters). In addition, Prevotella (phylum Bacteroidetes) shows a borderline significant elevation towards cluster 1. However, the other most abundant communities at genus level do not have clear differences across the biological clusters. As we observed in figure 6.7 on page 118, Streptococcus (phylum Firmicutes) was not statistically different between asthma and COPD at disease level, but significantly different across the identified biological clusters. This observation, suggested that the biological clusters can reveal some hidden patterns of the microbiome communities at genus level.

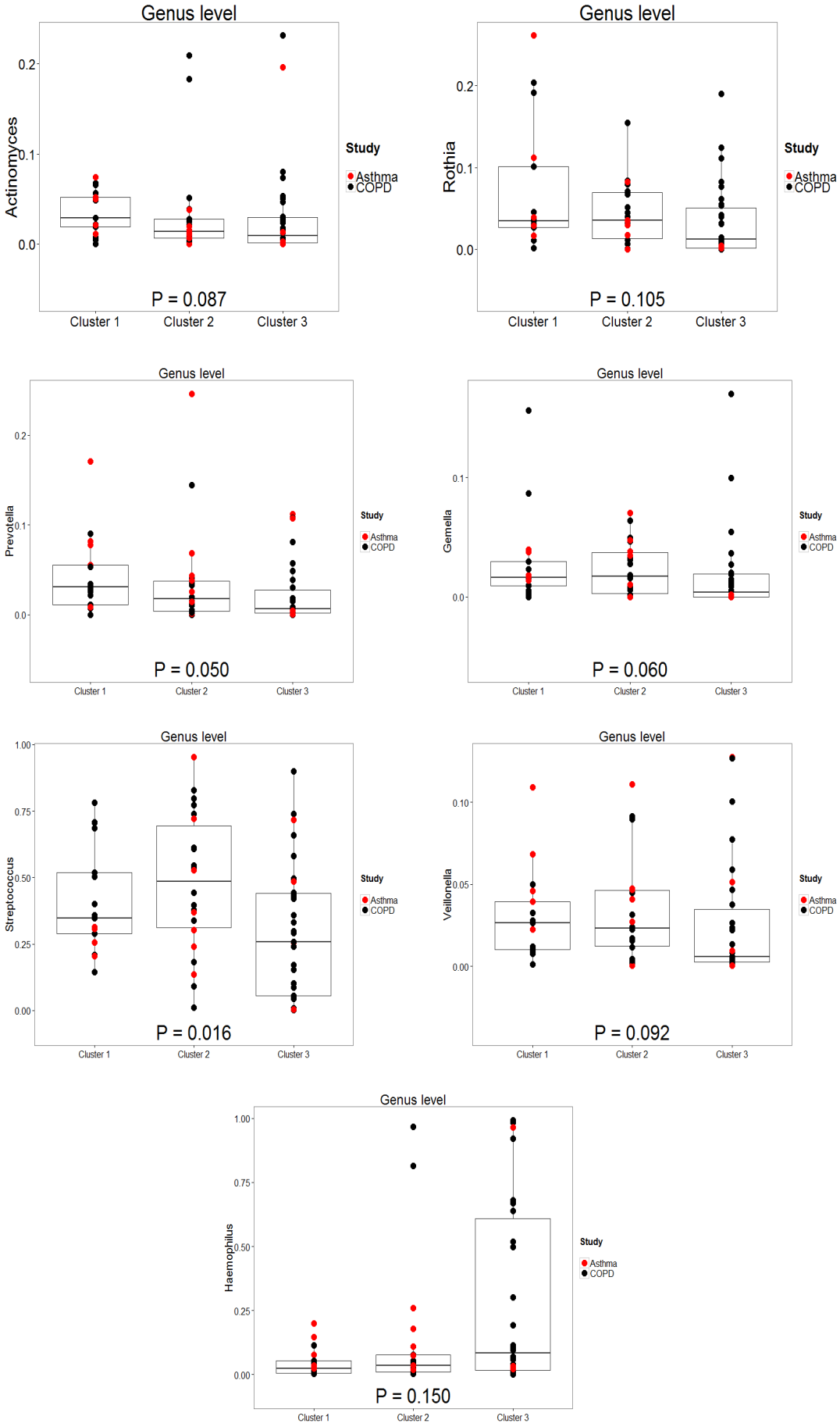


Figure 6.15: Pattern of the microbiome communities across the biological clusters at exacerbation

Summary of the Microbiome Communities Patterns

At phylum level

Alpha diversity is significantly higher in asthmatic, which suggests the proportions of the species in each subject are more diverse in asthmatic compared to COPD subjects. However, beta diversity (between subjects) is similar across the two diseases. In addition, both diversity are similar across the clusters.

From the most abundant species, only Bacteroidetes significantly elevated in asthmatic compared to COPD subjects at disease level, but it was not significantly different across the biological clusters. On the other hand, Firmicutes and Proteobacteria were not significantly different between the two diseases, but Firmicutes and Proteobacteria significantly elevated in cluster 2 and cluster 3, respectively.

At genus level

Alpha diversity is similar between the diseases, but Beta diversity is higher in COPD subjects compared to asthmatic, at genus level. In addition, Alpha diversity is not different between the clusters, but Beta diversity appeared to be larger in cluster 3 compared to clusters 1 & 2.

From these most abundant genera, only Prevotella (phylum Bacteroidetes) and Veillonella (phylum Firmicutes) are significantly elevated in asthmatics compared to COPD subjects at disease level; however, these species are not significantly different across the biological clusters. Whereas, although Streptococcus (phylum Firmicutes) was not significantly different across the diseases, but it appeared to elevate significantly in cluster 2 compared to the other clusters.

6.7 Discussion

In this study, three exacerbation biological subgroups of asthma and COPD were identified using the combination of factor and cluster analyses. The three subgroups represented both asthma and COPD with different proportion of overlap between the diseases. Subjects in cluster 1 have elevated eosinophils (blood and sputum) and T_H2 derived cytokines, and increased proportion (although not significant) of Actinobacteria and Bacteroidetes at phylum level, and Actinomyces (phylum Actinobacteria), and Prevotella (phylum Bacteroidetes) at genus level. Subjects in cluster 2 have high T_H1 derived mediators, and Firmicutes at phylum level, and Streptococcus (phylum Firmicutes) at genus level. Cluster 3 is a COPD predominated group which has high level in blood and sputum neutrophils, proinflammatory mediators, bacterial colonization and Proteobacteria at phylum level and sign of elevation in the proportion of Haemophilus (phylum Proteobacteria) at genus level although it was not significantly different across the clusters.

In addition, subjects in this study have records of clinical characteristics and sputum mediators at stable state, and their patterns were compared in each cluster with their corresponding exacerbation visits. In cluster 1, the lung function measurements such as pre- and post-FEV_1, pre- and post-FEV_1 percentage predicted were significantly lower at exacerbation compared to stable. In contrast, visual analogue scores (cough and dyspnea), and sputum eosinophils are significantly higher at exacerbation compared to stable state. In addition, most of the mediators have no significant difference between stable and exacerbation, except CCL-26 which is higher at exacerbation, in this cluster.

In cluster 2, the lung function measurements such as pre- and post-FEV_1, pre- and post-FEV_1 percentage predicted were significantly lower at exacerbation. In contrast, visual analogue scores (cough and dyspnea) increased at exacerbation, which is similar to the pattern observed in cluster 1. However, no blood or sputum cell-count was significantly different between exacerbation and stable states, which is dissimilar to the pattern in cluster 1. In addition, most of the mediators (such as IL-6, IL-10, IL-13, CXCL-10, CXCL-11, CCL-2, CCL-5 and TNFα) significantly increased but CCL-17 significantly decreased at exacerbation compared to stable state in this cluster.

In cluster 3, the lung function measurements (such as pre- and post-FEV_1, pre- and post FEV_1 percentage predicted) are significantly lower at exacerbation. On the other hand, visual analogue scores (cough and dyspnea), and blood and sputum neutrophil and sputum total-cell-count are significantly increased at exacerbation compared to stable

state. However, blood or sputum eosinophils are significantly lower at exacerbation compared to stable state in contrast to the patterns observed in cluster 1. In addition, the overall patterns of the mediators across stable and exacerbation were not quite clear. Mediators (such as IL-5, IL-6R and IL-10) were significantly higher at exacerbation; in contrast, most T_H1 and T_H2 mediators (such as IL-5, CXCL-11, CCL-2, CCL-13, CCL-17 and CCL-26) are significantly lower at exacerbation compared to their patterns at stable state.

Overall, each cluster has specific clinical and biological interpretation, and shows interesting patterns with respect to microbiome communities (at both phylum and genus levels). These clusters may represent a specific phenotype which may respond differently to particular treatment. In addition, the patterns across the stable and exacerbation subgroups are quite similar in which subjects who have elevated level of T_H2 cytokines have high eosinophil cell counts, and these which are high in proinflammatory mediators appeared to have high level in neutrophils cell counts and bacterial colonization. Furthermore, the two diseases appeared to have more in common in terms of subgroups at exacerbation compared to stable state.

The possible limitation of this study is that the asthmatic subjects at exacerbation are relatively smaller compared to COPD subjects. However, the subjects in each disease are reasonable and have enough power for any statistical and subgroup analyses. In addition, the exacerbation subgroups (unlike to the stable clusters) were not validated in an independent study as we do not have exacerbation validation dataset at the moment, so the stability of these clusters could be questionable and should be interpreted cautiously. Furthermore, as the number of subjects who have the microbiome information were relatively smaller than the subjects who have sputum cytokines (used for the identification of the biological clusters) so the entire pattern of the microbiome communities may not be reveal across the clusters. This may underpowered to establish a proper connection between the biological subgroups and the microbiome communities in this study and requires further studies with bigger sample size.

In conclusion, in this study three biological exacerbation subgroups were identified, which have specific clinical and biological interpretation. These three subgroups represented asthma and COPD with different proportion of overlap between both diseases.

Part II

Developing a Novel Method for Variable Selection in Model-based Clustering

Chapter 7

Variable Selection in Model-based clustering: a Finite Gaussian Mixture Model

7.1 Objectives

The main objective of this part of the thesis is to develop a new method of variable selection for model-based clustering (Gaussian mixture model).

7.2 Introduction

Although it seems the more information that exists for individuals would be better for clustering, adding non-informative (irrelevant) clustering variables may not make a basic change in the identification of the optimal clusters except hiding the existing subgroups. For example, assuming the observations are distinctive on few variables but similar (homogeneous) on most of the variables. Therefore, including these non-informative variables as input into the clustering algorithm could be a harmful (which may add unnecessary noise and hide the optimal clusters) as the clustering non-informative variables may dominate the effect of the clustering informative variables.

The general objective of variable selection in clustering is to maximize the identification of the optimal number of clusters using minimum number of clustering relevant variables. This may allow explaining the clusters in a simple and manageable way by removing these clustering irrelevant variables. In addition, it may improve interpretation, visualization, identification of cost-effective variables for future prediction and validation by not measuring those non-informative variables. Particularly in medical research it would be useful for subject selection for clinical trials, adjustment of the standard care treatments targeting each subgroup based on the state (threshold) of these relevant variables.

The application of clustering in biomedical research and other fields to the discovery of novel clusters/subgroups has increased considerably. However, selecting the best subset of relevant variables that produce optimal clusters remains a key problem. In respiratory research, to my knowledge, no single study applied a formal variable selection in clustering. The common approaches of choosing input variables for clustering algorithm are: based on the clinical utilities of the variables by eliminating the redundant ones [14], reducing the dimension of the observed variables using factor analysis [128]/ principal component analysis (PCA) [91] and use the corresponding scores as input into clustering algorithm, or using the highest-loading observed variables (after factor analysis) [87, 90].

These approaches have several limitations; for example, reducing variables based on their clinical utilities/previous information is very subjective without mathematical justification, in which the selected variables are not tested using rigorous mathematical techniques whether they represent well these excluded variables from the clustering algorithm. In addition, using factor/PCA scores in a situation where the observed variables don't have strong correlation is quite dangerous, as it is more likely to end up with few latent variables (scores) which represent very small amount of total information/variance exist in the entire variables as it has been observed in this study [91]; in which they reduced their variables into small PCA scores which only explain 61% of the information of the total variance of the observed variables. Furthermore, using the highest loading observed variables as input into clustering algorithm may not work well in a situation that were described in the simulation study (chapter 3). In summary:

1. The observed variables may not have strong correlation (internal patterns), and may not be suitable for factor analysis.

2. The factor loadings of the retained highest-loading observed variable may not be as close as to the optimal (i.e. to one).

3. There could be multiple observed variables that have similar loadings in each factor, but their loadings may not be as close as to the optimal (e.g. less than 0.7).

4. The entire information in the observed variables may not be captured reasonably well by the retained factors (corresponding to the highest-loading observed variables).

Using the above approaches is not always guaranteed for optimal outcome (clusters). Thus, a rigor investigation (screening) using appropriate techniques should be implemented in order to choose the input variables for clustering. In this study, we proposed a new

variable selection method for model-based clustering (which integrates variable selection and clustering simultaneously) to generalize the approach of Raftery and Dean (2006, JASA 101,168-178) [129]. This method is developed for any dataset which have continuous variables (which do not have strong correlation or internal patterns as sputum cytokines) as input into a model-based clustering algorithm such as clinical data (e.g. demographic and clinical characteristics) and/or other social science research data in which the number of observations should be greater than the number of variables. However, for dataset in which the variables that have strong correlation as sputum cytokines, the two stage approach (factor and cluster analyses) appeared as a best method to identify the optimal clusters in the simulation study (see chapter 3 for details).

This part of the thesis started with a general description of cluster analysis. The detailed mathematical formula of model-based clustering (Gaussian mixture model that is optimized using EM-algorithm), the effect of variable selection in clustering and the implementation of variable selection in model-based clustering are described. In addition, the algorithm of Raftery & Dean [129] was reviewed in detail. The detailed description of the proposed algorithm for variable selection in model-based clustering is reported. This part concluded by comparing the performance of the new method with other existing techniques using real and simulated dataset (with known cluster membership), and eventually it is applied to respiratory data in order to identify novel clusters.

7.3 Cluster Analysis

Cluster analysis is a technique that splits a heterogeneous population into more homogeneous subgroups (see figure 3.4 on page 63 in chapter 2 for graphical demonstration of clustering). Most of the existing clustering techniques rely on heuristic methods that are based on similarity or dissimilarity distance measures (such as k-means and hierarchical clustering algorithms). Although this approach is computationally feasible and available in most open-source and commercial statistical software, the main criticism is that there is no well accepted criteria for choosing the optimal number of clusters. An alternative method is a model-based clustering, in which a more formal statistical procedure can be implemented to choose the optimal clusters (mixture components) using likelihood approach. Thus, this part of thesis will focus on model-based clustering.

7.4 Model-based Clustering

7.4.1 Gaussian Mixture Model

A Gaussian mixture model is a model-based clustering technique in which each component probability distribution (mixture component) is corresponding to a cluster. It assumes that the observed data come from heterogeneous (two or more) populations instead of from a single (homogeneous) population, it works by modeling each of the sub-populations separately and the overall population as a mixture of these sub-populations [130, 131]. The optimal mixture components (clusters) and cluster memberships are estimated using maximum likelihood, which is optimized using EM algorithm [132].

Gaussian mixture model with k components is formulated as follows:

$$f(X) = \sum_{k=1}^{k} \omega_k f_k(x/\mu_k, \Sigma_k) \tag{7.1}$$

Where ω_k is the non-negative mixing coefficient, which sums to one, and represents the prior probability of an observation coming from component (cluster) k; and $f_k(x)$ is the density function of cluster k with mean μ_k, and variance–covariance matrix Σ_k.

7.4.2 Bivariate Gaussian Mixture Distribution

A bivariate distribution of Gaussian mixture of two components (clusters) with varying means and variance-covariance matrices is simulated and displayed in figure 7.1 on page 141. The data are simulated with $\omega_1 = 0.4$, $\mu_1 = [0, -2]$, $\Sigma_1 = [1.0, 0.4, 0.4, 1.0]$; $\omega_2 = 0.6$, $\mu_2 = [5, 3]$, $\Sigma_2 = [0.2, 0.6, 0.6, 0.3]$. This means that cluster 1 consists of 40% of the observations with mean of X_1 and X_2 equal to 0 and -2, respectively; and their standard deviation is the same which is 1, and the correlation between X_1 and X_2 is 0.4. Cluster 2 consists of 60% of the observations, in which X_1 has mean = 5 and standard deviation = 0.2; and X_2 has mean = 3 and standard deviation = 0.3, the correlation between X_1 and X_2 (in this cluster) is 0.6. These data represented the general format of Gaussian mixture, and can be extended to high-dimensional multivariate Gaussian mixture distribution by increasing the number of variables.

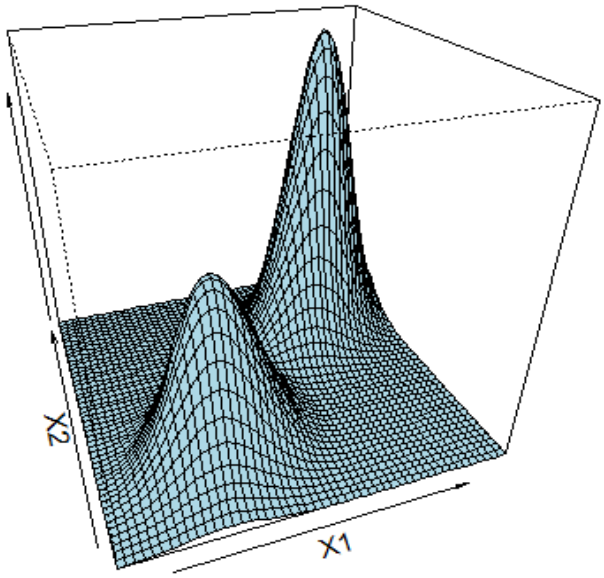


Figure 7.1: Bivariate Gaussian mixture density

7.4.3 Maximum Likelihood Estimation for Gaussian Distribution

Likelihood can be used to measure the goodness of fit in a Gaussian model, such as how well the model fits the data.

Assuming there is an independent dataset, $X = [x_1, x_2, ..., x_n]$, drawn from a single Gaussian distribution with probability density function $f(X, \theta)$; where θ [μ (mean) and σ^2 (variance)] are the parameters. The likelihood function can be written as follows:

$$L(X; \mu, \sigma^2) = \prod_{i=1}^{n} f(X_n; \mu, \sigma^2) \tag{7.2}$$

The goal is to estimate the parameters ($\hat{\mu}$ and $\hat{\sigma}^2$) that maximized the likelihood function (equation 7.2). However, it is impossible to maximise directly from the above equation. Therefore, the equation should be transformed to the corresponding natural logarithm (log-likelihood function), which is strictly equivalent to the likelihood function. Then the parameters can be analytically estimated by partial differentiating the log-likelihood function, with respect to the mean (μ) to get sample mean ($\hat{\mu}$), and with respect to the variance (σ^2) to get sample variance ($\hat{\sigma}^2$). This approach works well in a single Gaussian distribution to identify parameters which maximize the log-likelihood function. However, in multivariate Gaussian mixture model, it is difficult to implement this approach to estimate the parameters that maximise the corresponding log-likelihood

function (equation 7.3). Thus, an iterative approach such as EM-algorithm is used to estimate the parameters that maximize the log-likelihood function in multivariate Gaussian mixture model [132].

The log-likelihood function of Gaussian mixture model can be written as follows;

$$log\, L(\theta) = \sum_{j=1}^{n} log\{\sum_{k=1}^{K} \omega_k\, f(x_j; \mu_k, \Sigma_k)\} \tag{7.3}$$

7.4.4 EM-algorithm for Gaussian Mixture Model

EM is an Expectation - Maximization algorithm which maximizes the log-likelihood function (equation 7.3) of a Gaussian mixture model (equation 7.1) with respect to the estimated parameters (ω_k, μ_k, Σ_k) using an iterative approach [132]. The parameters ($\omega_k^{(0)}$, $\mu_k^{(0)}$, $\Sigma_k^{(0)}$) can be initialized using heuristic approach such as k-means algorithm. In the EM-algorithm, an observation's class membership is identified using Bay's theorem, by calculating the posterior probabilities, and assigning the observation into the group/cluster in which the observation has the highest probability.

Given a Gaussian mixture model (equation 7.1) with log-likelihood function (equation 7.3), the optimal parameters which maximize the log-likelihood function can be estimated by iterating the following *E and M steps* until convergence to the maximum likelihood function, i.e. $(log\, L(\theta^{i+1}) - log\, L(\theta^i) \approxeq 0)$.

E-step: Estimate the posterior probabilities using the current parameters

$$\tau_{jk}^{(i)} = \frac{\omega_k^{(i)} f(x_j; \mu_k^{(i)}, \Sigma_k^{(i)})}{\sum_{k=1}^{K} \omega_k^{(i)} f(x_j; \mu_k^{(i)}, \Sigma_k^{(i)})}$$

M-step: Re-estimate the parameters using the current posterior probabilities

$$\begin{aligned}
\omega_k^{(i+1)} &= \frac{\sum_{j=1}^{n} \tau_{jk}^{(i)}}{n} \\
\mu_k^{(i+1)} &= \frac{\sum_{j=1}^{n} \tau_{jk}^{(i)} x_j}{\sum_{j=1}^{n} \tau_{jk}^{(i)}} \\
\Sigma_k^{(i+1)} &= \frac{\sum_{j=1}^{n} \tau_{jk}^{(i)} (x_j - \mu_k^{(i+1)})(x_j - \mu_k^{(i+1)})^T}{\sum_{j=1}^{n} \tau_{jk}^{(i)}}
\end{aligned} \tag{7.4}$$

Here, the likelihood function is guaranteed to increase in each iteration until convergence. However, the convergence is not guaranteed to be a global maximum as it could converge to a local maximum, in which more than a single optimal result could be identified. In addition, the EM algorithm is prone to initialization in which the results could be influenced how the parameters were initialized, usually heuristic clustering algorithm such

as k-means [105] are commonly used to initialize the parameters. For more detailed proof, application of EM - algorithm to Gaussian mixture model and its strength and limitations, readers refer to GJ. Titterington et al (1985) [130] and GJ McLachlan et al (2007) [131].

7.4.5 Optimal Number of Clusters in Gaussian Mixture Model

The likelihood function of the Gaussian mixture may be used to assess how well the model fitted to the data. However, the corresponding maximum log-likelihood estimator gets larger as the number of components (clusters) is increased in the model. Therefore, the maximum log-likelihood estimator is not the suitable approach to choose the optimal k components/clusters.

However, the most common approach to address this problem in Gaussian mixture model is using the Bayesian information criterion (BIC). BIC was originally developed by G. Schwarz (1978) [133] for assessing non-nested model fit. Fraley and Raftery(1998) [134] showed that it is approximately equivalent to Bayes factor [135, 136], and they successfully applied for model comparison and choosing the optimal number of clusters in Gaussian mixture model [134]. It is insensitive to the number of clusters, and the smaller the BIC is the better fit.

BIC for Gaussian mixture model is formulated as follows:

$$BIC = -2 * log\, L(\theta) + p * log(n) \tag{7.5}$$

Where, $L(\theta)$ is the maximum log-likelihood estimator, n is the number of observations; $p = \frac{2*c-2+3*c*d+c*d^2}{2}$ (number of parameter), c= number of clusters, and d= number of variables.

7.5 Variable Selection in Cluster Analysis

7.5.1 Introduction

Variable (feature) selection is quite common in supervised techniques such as linear regression and discriminant analyses [137–139]. However, recently, with the availability of hundreds of thousands of variables, such as gene expression and medical imaging data, some progress has been made for formalising the objective of variable selection in unsu-

pervised techniques such as cluster analysis, and that become an active field of research [129, 140, 141].

Although it seems more information that exists for individuals would be better for clustering, but adding non-informative (irrelevant) clustering variables may not make a basic change in the identification of the optimal clusters except for hiding the existing subgroups. For example, assuming the observations are distinctive on few variables but similar (homogeneous) on most of the variables. Therefore, including these non-informative variables as input into the clustering algorithm could be a harmful (which may add unnecessary noise and hide the optimal clusters) as the clustering non-informative variables may dominate the effect of the clustering informative variables.

The general objective of variable selection in clustering is to maximize the identification of the optimal number of clusters using minimum number of clustering relevant variables. This may allow the explanation of the clusters in a simple and manageable way by removing these clustering irrelevant variables. In addition, it may improve interpretation, visualization, identification of cost-effective variables for future prediction and validation by not measuring those non-informative variables [139].

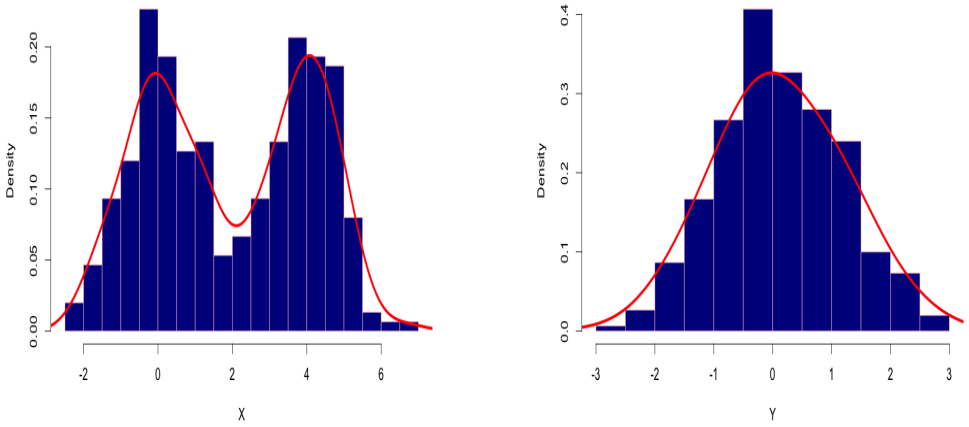
There are two common types of variable selection in clustering. They are filter and wrapper approaches. In the filter approach, variables are selected prior to clustering [142, 143] based on certain criteria (e.g. based on previous studies or data reduction techniques such as principal component or factor analyses). It is computationally feasible; however, there is a risk to remove potential clustering informative variables through that filtering process. Whereas, the wrapper method is an alternative approach (which address the issue in the filter approach) that assesses the variables according to their clustering usefulness, in which variable selection and clustering are implemented simultaneously [129, 140, 144]; in which this part of the thesis will focus on.

7.6 Variable Selection for Model-based Clustering

7.6.1 Clustering Relevant and Irrelevant Variables

In cluster analysis, variables can be categorized as clustering relevant (informative) or irrelevant (non-informative) based on their ability to split the observations into significant distinctive subgroups. A variable that can classify observations into distinctive subgroups/clusters is called a clustering relevant variable, if this is not the case it is known as a clustering irrelevant variable. For example, the simulated variable X in figure 7.2(a)

can be treated as a clustering relevant variable as it is able to split the observations into two distinctive subgroups at approximate cutoff 2 (e.g. at $X = 2$). However, variable Y in figure 7.2(b) can be considered as clustering irrelevant as it could not split the observations into distinctive subgroups. Although the observations can be classified into two subgroups on the bases of variable Y (e.g. at $Y = 0$), the difference between the two subgroups would more likely be insignificant (i.e. homogeneous across the subgroups on Y). Thus, including variable Y into a clustering algorithm does not add any further information except for including unnecessary noise which could hide the existing clusters with respect to variable X.



(a) (b)

Figure 7.2: (a) Clustering relevant and (b) Clustering irrelevant

7.6.2 Univariate Clustering Variable Selection

In model-based clustering a variable relevance for clustering can be assessed univariately using BIC index.

For example, by calculating the BIC-differnce (ΔBIC) between "no-cluster" to "optimal-clusters", and formulated as follows:

$\Delta BIC = BIC_{noCluster} - BIC_{optimalClusters}$

The standard calibration of BIC difference with respect to clustering evidence is reported in [134, 135], and formulated as follows:

1. $BIC_d < 2:$ Weak evidence
2. $BIC_d \geq 2 \,\&\, BIC_d < 6:$ Positive evidence.
3. $BIC_d \geq 6 \,\&\, BIC_d < 10:$ Strong evidence
4. $BIC_d \geq 10:$ Very strong evidence

Example

Several normally distributed univariate data were simulated from two subgroups, which have varying means but (for simplicity) the same standard deviation, and displayed in figure 7.3. Then BIC for no-clusters (i.e. one cluster) and for two-clusters (optimal clusters) were estimated as formulated in equation 7.5 on page 144, and the BIC differences were calculated.

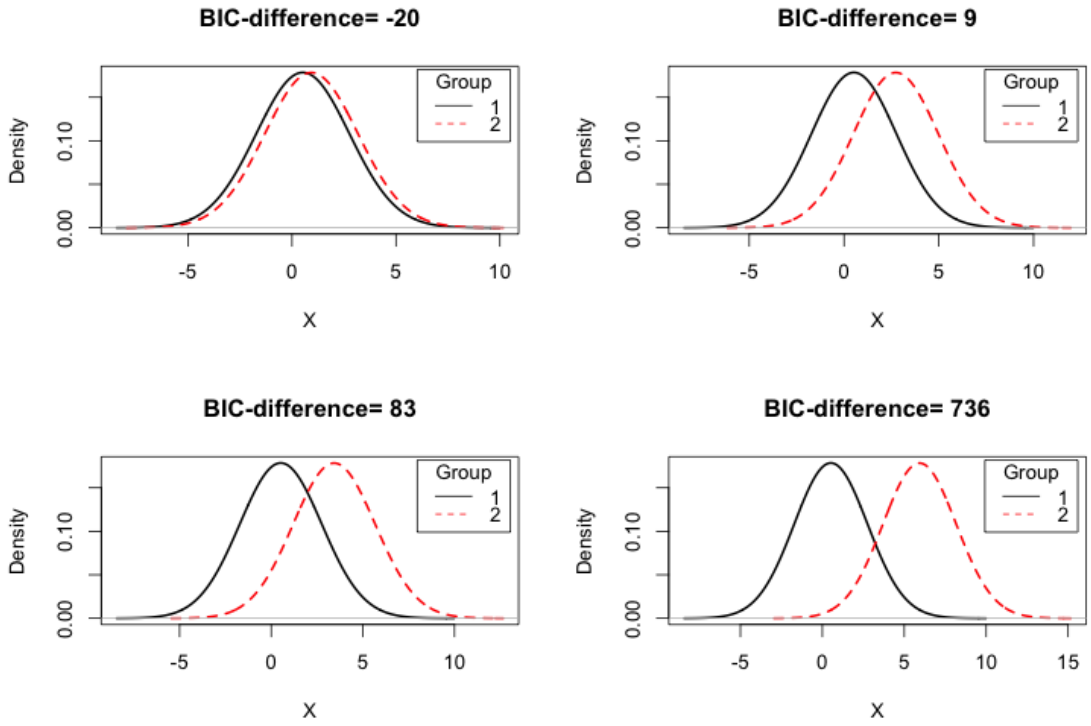


Figure 7.3: BIC for assessing clustering information

As we observe in the figures above, when the mean difference between the two groups/clusters is very small (almost zero) the BIC difference between no-cluster and two-clusters becomes negative, which suggests that the observations should be treated as one group/cluster instead of two subgroups. However, as the mean differences between the two groups increases, the corresponding BIC differences appeared to increase substantially. Thus, BIC is a useful criterion for assessing a variable for clustering relevance.

7.6.3 Multivariate Clustering Variable Selection

As the dimension of the variables increases, selecting clustering relevant variables is not straightforward as the relevance of a variable may be influenced by other variables' structure. In general we expect four possible structures (scenarios) in high dimensional data. For graphical demonstration, an artificial data was simulated (which fairly represented the four possible scenarios) and displayed in figure 7.4 on page 150.

Scenarios

1. Clustering relevant and correlated variables

    - For instance, two variables which are correlated (dependent) and both have clustering information (see figure 7.4(a)).

2. Clustering relevant and irrelevant but uncorrelated variables.

    - For example, two variables in which one is clustering relevant and the other is irrelevant, and they are uncorrelated (see figure 7.4(b)).

3. Two clustering relevant and but uncorrelated variables

    - For instance, one variable which is clustering relevant and the other one is irrelevant but they are dependent (correlated) (see figure 7.4(c)).

4. Clustering relevant and irrelevant but correlated variables

    - For example, one variable which is clustering relevant and the other one irrelevant but they are uncorrelated (independent) (see figure 7.4(d)).

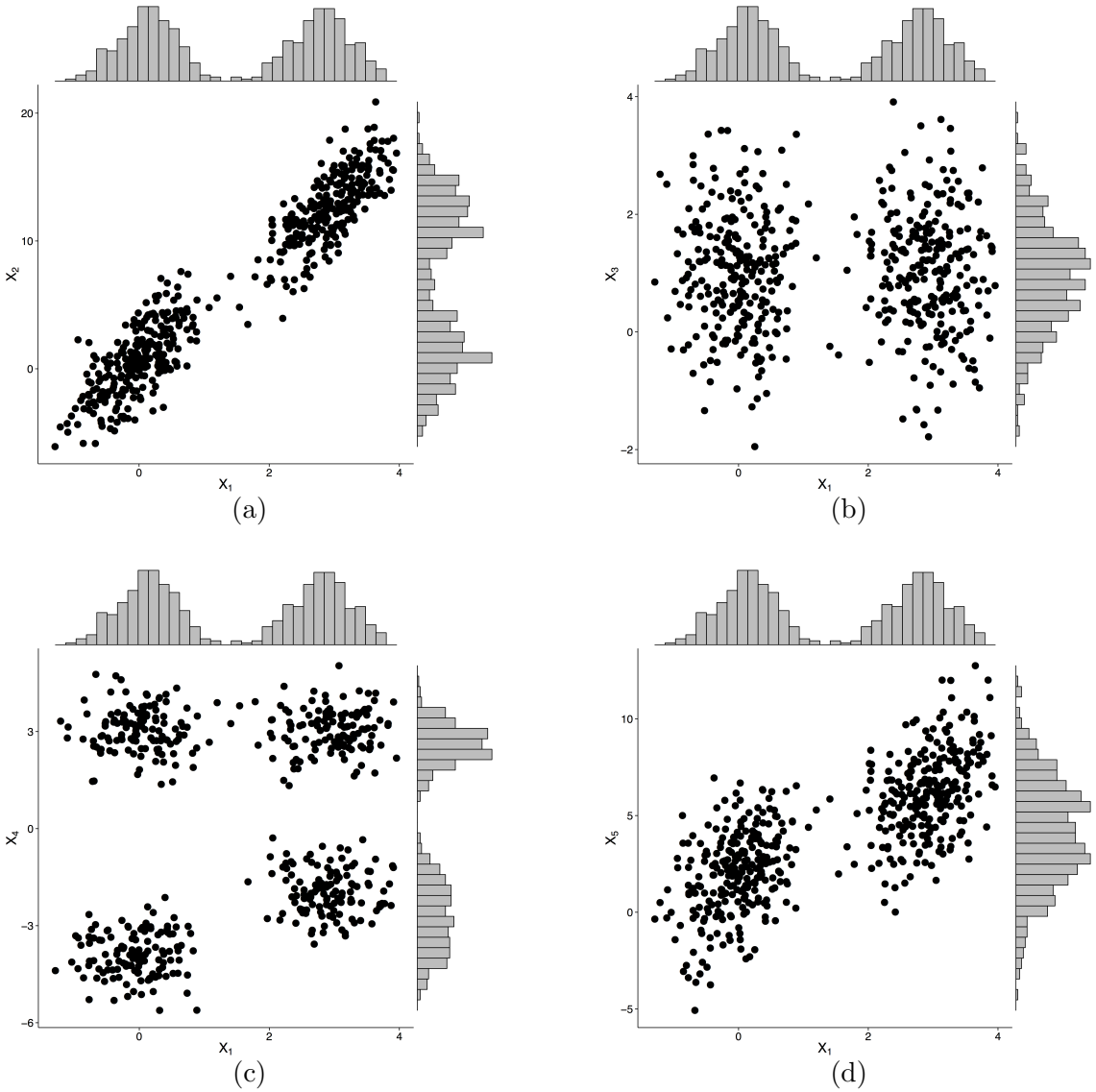


Figure 7.4: (a) Scenario 1: Both clustering informative and dependent variables; (b) Scenario 2: Clustering informative and uninformative but uncorrelated variables; (c) Scenario 3: Both clustering informative but independent variables; (d) Scenario 4: Clustering informative and uninformative but dependent variables

Thus a robust method, which accounts the above possible scenarios, should be implemented in order to identify these clustering relevant variables in high-dimensional (multivariate) data.

7.6.4 Regression-based Variable Selection

Variable selection in clustering is very complex particularly as the dimension of the variables increases, because the relevance of a variable is considerably affected by its correlation with other relevant or irrelevant variables. However, Raftery and Dean (R&D)[129] attempted to address this challenge as a model comparison problem using forward stepwise variables selection algorithm. They evaluated the dependence between the relevant and irrelevant variables using linear regression. In other words, they assessed how well the irrelevant variable was represented by those relevant variables, by regressing the clustering irrelevant variables on the relevant variables and calculated the corresponding regression BIC.

Their paper raised motivation for several researchers (e.g. Maugis C and colleague [140] attempted to extend their algorithm); however, its limitation is not yet fully addressed. Inspiring by their method, I started to develop a new variable selection method for model–based clustering, which extends their approach. The variable selection in the R&D method entirely depends on the self-standing model-based clustering technique [145, 146] (in which the EM-algorithm was initialized using hierarchical algorithm). However, in this study, I wrote my own new mode-based clustering technique, in which the EM-algorithm was initialized using several heuristic algorithms such as k-means, kmediod or fuzzy k-means. In addition, a new approach also proposed to address the singularity issues in the model-based clustering.

Overview of Raftery and Dean Method

Raftery and Dean (2006, JASA 101, 168-178) [129] developed a regression-based variable selection method for model-based clustering

The basic idea of the R&D algorithm is as follows:

- Assume a dataset has X_1 and X_2 measurements, and X_1 has more clustering evidence than X_2 based on BIC.

- **Step 1:** Evaluate optimal clustering BIC for X_1 (BIC_{X_1})
- **Step 2:** Regress X_2 on X_1 and calculate regression BIC (BIC_{reg}).
- **Step 3:** Identify optimal clusters using combination of X_1 and X_2, and record the joint BIC (BIC_{joint}).
- **Step 4:** If $BIC_{joint} < (BIC_{X_1} + BIC_{reg})$, then X_2 will be included as clustering relevant with X_1, otherwise it will be excluded as irrelevant.

Regression BIC (BIC_{reg}) is calculated as follows:

$$BIC_{reg} = -n\,log(2\pi) - n\,log(RSS/n) \; - \; n - (dim(S^1) + 2)\,log(n) \tag{7.6}$$

Where, RSS is the residual sum of squares; n is the number of observations; $dim(S^1)$ is the dimension (number) of relevant variables

The R&D method performs reasonable well in selecting the clustering relevant variables in real and simulated data. However, as the structure of the dataset gets so complex, it appears to miss some clustering relevant variables and as a consequence the expected number of optimal clusters is also lost.

The following may be the possible reasons for their method limitations;

1. They considered unnecessary relationship between clustering relevant and irrelevant variables; and did not allow the irrelevant variables to be independent of the clustering relevant variables in a situation where the two variables are uncorrelated.

2. They considered only linear relationship between the clustering irrelevant and relevant variables although there is a possibility that the relationship could be non-linear, which may affect the regression BIC.

3. To avoid the singularity of the log-likelihood function (i.e. variance-covariance matrix is not positive definite), they constrained the variance-covariance matrix, which may have a substantial impact in their algorithm to miss the optimal clusters and misclassify the observations.

Motivating Example

Multivariate data with known clusters was simulated. This data has six clusters and 8 clustering relevant and irrelevant variables (X_1, X_2, ..., X_8). Variables X_1, X_2, X_3, X_4 and X_5 are clustering relevant, and X_6, X_7 and X_8 are clustering irrelevant variables (homogeneous across the clusters). However, variables X_1, X_2 and X_3 are uncorrelated to the other clustering relevant variables (X_4 and X_5). The data is displayed in figure 7.5 on page 154.

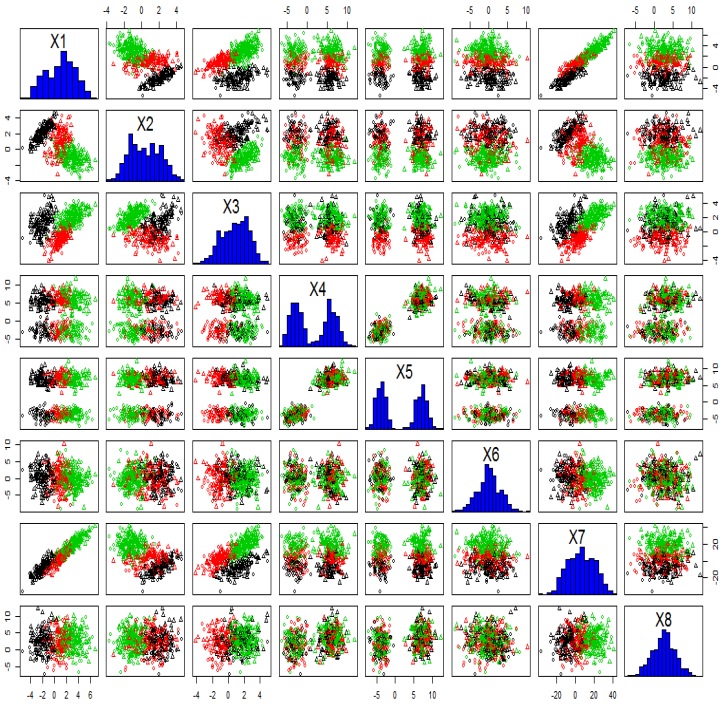


Figure 7.5: Multivariate simulated data, the combination of colours (red, black or blue) and symbols (circle or triangle) divided the data into six subgroups.

The performance of the R&D method was assessed on the simulated data above (which is displayed in figure 7.5). Thus, their method selected only X_4 and X_5 as clustering relevant variables, and two corresponding optimal clusters (which are much smaller than the expected 6 simulated clusters).

Scenarios to Address This Problem

Two approaches were proposed to address that particular problem which was observed when the R&D method was applied to the above simulated data.

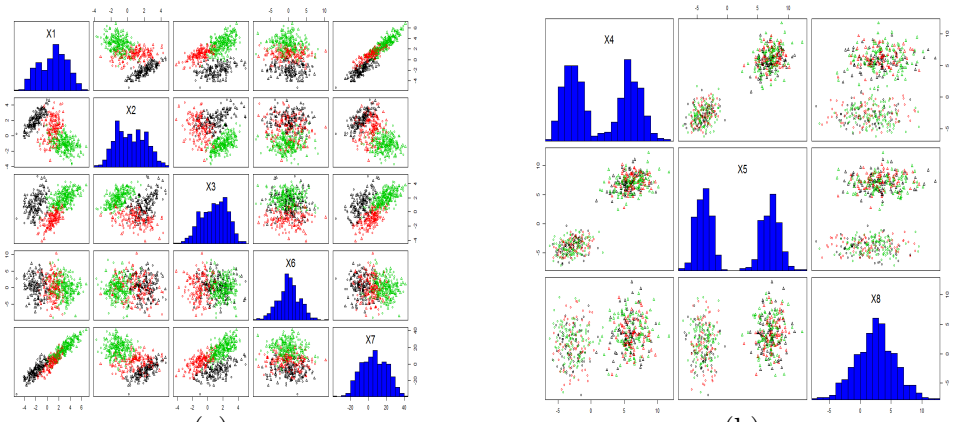
First Scenario:

In multivariate data, variables may have hidden internal structures (based on their correlations), and on the bases of these patterns variables can be partitioned into independent subsets. Variables which are strongly correlated can be assigned into the same subgroup. Then the variable selection method can be implemented in each variables' subset in order to identify the corresponding clustering relevant variables.

Steps for the Proposed Scenario

- Firstly, the entire variables are partitioned into independent subsets based on their correlation matrix (internal structures)
- Variables which are strongly correlated are assigned together
- Clustering relevant variables are selected from each subset
- Finally, using the aggregated relevant variables from each subset, the corresponding optimal clusters are identified.

The above proposed approach was applied to the simulated data which is displayed in figure 7.5, and some hidden structures were identified within the variables. Thus, based on those structures the variables were partitioned into two independent subsets, and are displayed in figure 7.6 (a) & (b) on page 155. Variables X_1, X_2, X_3, X_6 and X_7 were assigned together, and variables X_4, X_5 and X_8 grouped together as second subset.



(a) (b)

Figure 7.6: (a) First subset and (b) Second subset. The combination of colors (red, black or blue) and symbols (circle or triangle) divided the data into six subgroups.

By utilizing the R&D method into each variables' subgroup, variables X_1, X_2 and X_3 from the first subset, and variables X_4 and X_5 from the second subset were selected as clustering relevant variables. Then using these five relevant variables (X_1, X_2, X_3, X_4 and X_5), six optimal clusters (as expected) were identified using mclust package [145] with misclassification of 15% .

This shows that how the identification of clustering relevant variables could be improved just by partitioning the entire variables into independent subsets based on their

correlation patterns (internal structures). However, the mclust package [145] assigned only 85% of the observations into the true clusters (although it identified the optimal number of clusters), which needs some modification for optimal observations‘ assignment into the right clusters.

Second Scenario

There is a scenario (situation) in which variables may not have a relationship with respect to the entire (global) dataset, but a relationship (correlation) may exist among the variables locally in part of the data (e.g. in each cluster). To assess for this pattern (non-linear relationship) between the irrelevant and relevant variables, the following additional approach was proposed.

- Once the global relevant variables and corresponding optimal clusters are identified using the first scenario.
- Each cluster is treated as an independent dataset, and a further search in each cluster for local relevant variables from these globally irrelevant variables is performed.

To demonstrate the second scenario graphically, a bivariate data is simulated and depicted in figure 7.7 on page 157. Assume that only X was selected as clustering relevant variable and two corresponding optimal clusters were identified using the first scenario. However, if each cluster was assessed further with respect to Y, then additional local clusters appeared within each global cluster. Then using both X & Y, four optimal clusters could be identified. Thus to avoid loss of such hidden important clusters, the second scenario needs to be accounted for.

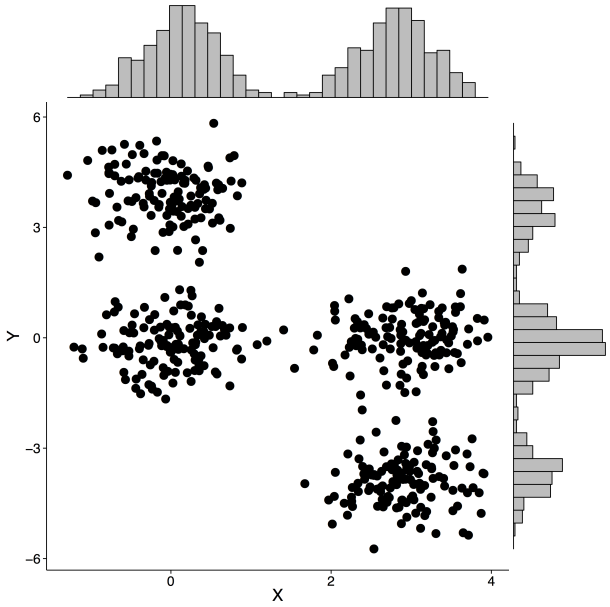


Figure 7.7: Additional clusters within a cluster

In the above multivariate simulated data 7.5 on page 154, the six global clusters were treated as new six independent datasets. Then the global irrelevant variables (X_6, X_7 and X_8) were further examined in each cluster whether they could be able to classify the clusters into further subgroups. However, no additional clusters were found apart from those previously identified six global clusters using the first scenario.

7.7 Proposed Variable Selection Method for Model-based Clustering

To formalize the above scenarios into an algorithm, we propose a new variable selection method for model-based clustering which generalizes the approach of R&D. This method relaxes the global prior assumptions of linear-relationships between relevant and irrelevant variables by searching for latent (hidden) structures among the variables; and also accounts for non-linear relationships between relevant and irrelevant variables. A Gaussian mixture model (with unconstrained variance-covariance matrices) was fitted using the EM-algorithm (equations 7.4.4 and 7.4 on page 143), and was used to identify the optimal clusters.

7.7.1 Initialization the Parameters in the EM-algorithm

As the EM-algorithm is prone to initialization in which it could converge to a local maxima rather than the global maximum, in such a situation there may be multiple optimal solutions (clusters) for a single dataset, especially when the clusters are not well separated. Thus, in this study the parameters in the EM-algorithm were initialized using several heuristic approaches such as the k-means [105], k-medoids, fuzzy-kmeans or hierarchical; in which researchers could have an option to choose which could suit to their data structure.

7.7.2 Covariance Matrix Singularity

In the application of EM-algorithm to the Gaussian mixture model, singularities (variance-covariance not positive definite) in the likelihood function is quite common particularly when the number of data in a cluster becomes insufficient. For example, when a component collapses onto a data point (this means that where a cluster or component has only a single observation) in which the mean is equivalent to the data-point and the variance is zero, and the corresponding log-likelihood becomes indefinite (infinite) (see eq 7.3 on page 142 for mathematical details).

There are various approaches in the literature to keep the variance-covariance matrix positive definite (avoiding singularity). For example, R&D variable selection method used a self-standing model-based clustering technique [146] in which the problem was addressed based on the eigenvalue decomposition of the mixture components variance-covariance matrices. They constrained the variance-covariance matrix, and identified 14 possible models to fit the data which optimize the variance-covariance matrices that ensures a positive definiteness in at least one of the models. Although that approach guarantees for positive definite, but not necessary is the best fit to the data (here is a danger in which that may not reveal the optimal number of clusters and may not also assign the observations into the expected right subgroup).

In this study, we use the Gaussian mixture model with unconstrained (full) variance-covariance matrices (unlike R&D method) which accounts various dependencies among the variables within and across the clusters. For example, variables may be dependent in one cluster but independent within another cluster. We handled the singularity issues using the algorithm proposed here [147]. The algorithm is written in R statistical software and reported in appendix A.

The singularity algorithm is described below:

1. First the variance-covariance matrix is tested for positive definiteness using Cholesky factorization
2. While the test fails, one of the following is done:
   (a) If there is a negative element in the diagonal, then one percent of the maximum element in the diagonal is added to all diagonal elements
   (b) Otherwise all the diagonal elements are incremented by one percent.
3. Iterate step 1 and 2 till the variance-covariance matrix have positive definite.

7.7.3 Application of Factor Analysis to Split Variables into Independent Subsets

In this study factor analysis was used for assessing the hidden internal structures and splitting of the variables into independent subsets. First, factor analysis with varimax rotation (see section 3.3.3 on page 59 for details) is performed to the entire dataset, and factors which have eigenvalue above one are retained, then variables are assigned (subset) together if they load in the same factor, otherwise partitioned into different subset. For example, in the simulation study in chapter 3, factor analysis with varimax rotation was performed to the artificial data, and variables X_1, X_3, X_{13}, X_{14}, & X_{18} loaded in factor 1; variables X_2, X_8, X_{10}, X_{16}, & X_{17} in factor 2; variables X_6, X_7, X_9, X_{19}, & X_{20} in factor 3; and variables X_4, X_5, X_{11}, X_{12}, & X_{15} in factor 4 (see table 3.2 on page 62 for details). Therefore, in the proposed variable selection for model-based clustering method, variables which load in the same factor (after factor analysis with varimax rotation) are assigned together in the same subset (variables' subgroup).

7.7.4 Proposed Algorithm

The fundamental basis of the proposed method is to start with one large variables set containing the entire variables, and then divide the variables into independent subsets based on their correlation-matrix. Variables that are dissimilar (uncorrelated) are split off and turned into small independent subsets. The forward stepwise variable selection algorithm is applied separately on each variables' subset and split the variables into clustering relevant and irrelevant. Then the global relevant variables are aggregated from each subset, and used to identify the corresponding optimal clusters. In addition, each cluster is considered as independent new dataset, and a further search for local relevant variables from these global irrelevant variables is performed. Finally, the relevant variables are updated, and the corresponding optimal clusters are identified.

The proposed approach accounts for the possible structures of variables in high-dimensional data whether they are correlated or uncorrelated by searching the latent (hidden) patterns of the variables. In addition, it accounts for the linear and non-linear relationship between the relevant and irrelevant variables. The proposed method comprises three algorithms.

1. Algorithm 1: For splitting the variables into independent subsets

2. Algorithm 2: For forward stepwise variable selection in each subset.

3. Algorithm 3: For implementing algorithms 1 & 2 further in each cluster

Proposed Algorithms

Algorithm 1

Step 1: *Let Y denote the entire variables. Apply factor analysis with varimax rotation to Y.*

Step 2: *Retain factors having eigenvalue greater or equal to one.*

if $(N(F) > 1)$ **then**

Step 3: *Identify a factor in which a variable has maximum loadings*

Step 4: *Subset variables together if they have maximum loadings in the same factor*

else

Stop

end

Algorithm 1: For splitting the variables into independent subsets. $N(F)$ = number of factors having eigenvalue greater than one.

Algorithm 2

Step 1: *Let S denote one of the variables subset. Apply model-based clustering to each variable in S. Choose the variable yielding the highest BIC difference (ΔBIC) between optimal-clusters and no-clusters, and record ΔBIC and optimal $-$ BIC for that variable, S^1.*

if $(\Delta BIC > 0)$ **then**

- ***Step 2:*** *Split the variables (S) into clustering relevant (S^1) and irrelevant (S^2)*
- **while** $(N(S^2) > 0)$ **do**
  - ***Step 3:*** *Fit linear regression for each variable in S^2 on S^1 and calculate corresponding regression BIC (BIC_{reg}), and record summation of S^1 optimal-BIC (BIC_{opt}) and BIC_{reg} for each variable (X_i) in S^2 and denote as $sumBIC = BIC_{opt} + BIC_{reg}$.*
  - ***Step 4:*** *Apply model based clustering on the combination of S^1 and each variable from S^2, and calculate the joint optimal-BIC (BIC_{joint}), and choose the variable which has the highest BIC difference ($BIC_{diff} = sumBIC - BIC_{joint}$)*
  - **if** $(BIC_{diff} > 0)$ **then**
    - Include that variable with the relevant variables (S^1)
    - Repeat ***steps 3 and 4*** *till no new variable is added to S^1*
  - **else**
    - *Stop*
  - **end**
- **end**

else

- *Stop*

end

Algorithm 2: For forward stepwise variable selection in each variables' subset.

Algorithm 3

while $(N(S^1) > 0 \quad \& \quad N(S^2) > 0)$ **do**

Step 1: *Apply model-based clustering on* S^1 *and record the optimal number of clusters (*C*).*

Step 2: *Treat each global cluster* j (C_j)*, as separate new dataset*

Step 3: Search for local relevant variable in each *cluster*(C_j) on S^2 using algorithms 1 & 2

if $(N(S^1_{c_j}) > 0)$ **then**

Step 4: *Include these variables with* S^1

else

Stop

end

end

Algorithm 3: For implementing algorithms 1 & 2 further in each cluster. $N(S^1)$ = number of relevant variables; $N(S^2)$ = number of irrelevant variables; $N(S^1_{c_j})$ = number of relevant variables in each cluster.

7.8 Performance of the Proposed Method

The proposed algorithm is written in R statistical software [113], and developed into R-package ("VarSel4GMM"), the codes are reported in appendix B. In the proposed method, the Gaussian mixture model with unconstrained (full) variance-covariance was used, which was optimized using EM-algorithm. The EM-algorithm was initialized using several heuristic clustering algorithms such as k-means, k-medoids or fuzzy k-means. To assess the performance of the proposed method, simulated and real datasets with known class membership were used; and its performance was compared with the R&D method as their algorithm is available in R-package ("clustvarsel").

7.8.1 Instructions on How to Use the "VarSel4GMM" Package of the Proposed Method in R

The R package ("VarSel4GMM") for the proposed method comprises six interlinked R functions such as "gmmEM","sigmaFixer", "REGbic", "EMvSel", "subEMvSel" and "gmmVarSel", and are presented in appendix A. In this section, how to use the package (in R scientific computing platform) for simultaneous variable selection and clustering is described.

gmmEM

This function implements the model-based clustering (Gaussian mixture model which is optimized using EM algorithm) to the dataset. The EM-algorithm is initialized using k-means, k-medoids or fuzzy k-means. The input is a dataframe or matrix (rows = observations and columns=variables). The outputs are number of clusters, assignment of observations in each cluster, means and variance-covariance matrix of input variables in each cluster, Bayesian information criterion (BIC), maximum log-likelihood estimator, number of iteration (after how many iterations the algorithm converges or stops), proportion of observations assigned in each cluster and number of parameters (degree of freedom).

sigmaFixer

This function fixes the singularity issues in variance-covariance matrix of Gaussian mixture model (which is optimized using EM-algorithm). First, it computes the Choleski factorization of the variance-covariance matrix to assess for positive definite; then if the matrix

is not positive definite, it keeps adding 1% to the diagonal of the matrix and iterates till the matrix is positive definite.

REGbic

This function calculates the univariate and multivariate regression BIC. The dependent variable is a single continuous variable but the predictors can be a single or multiple continuous variable/s.

EMvSel

This function implements the forward variable selection algorithm to the dataset in order to identify the clustering relevant variables in model-based clustering.

subEMvSel

This function splits the variables into independent subsets and implements the "EMvSel" function in each variables' subset. Then the global clustering relevant variables from each subset are selected and the corresponding global optimal clusters are identified.

gmmVarSel

This function treats the global clusters, which are identified using the "subEMvSel" function, as independent new datasets and further search for local relevant variables from these globally irrelevant variables is implemented in each cluster, and returns the final updated clustering relevant variables. Then these variables are plugged into the "gmmEM" function and the final corresponding optimal clusters are identified.

7.8.2 Example 1: Simulated data

First simulated data

An artificial data on eight dimensions was simulated, which consists of two subsets of clustering relevant variables (X_1, X_2 & X_3) and (X_4 & X_5) (with no correlation between the two relevant subsets), and normally distributed clustering irrelevant variables (X_6, X_7 and X_8). The first relevant subset (X_1, X_2 and X_3) divided the entire dataset into three distinctive subgroups, and the second subset (X_4 and X_5) into two subgroups. From these irrelevant variables, X_6 is a noisy variable which does not correlated with any of the variables, but X_7 only correlated with the first relevant subset (X_1, X_2, and X_3) and X_8 correlated only with the second relevant variables (X_4 and X_5). The data is displayed in figure 7.8 on page 167.

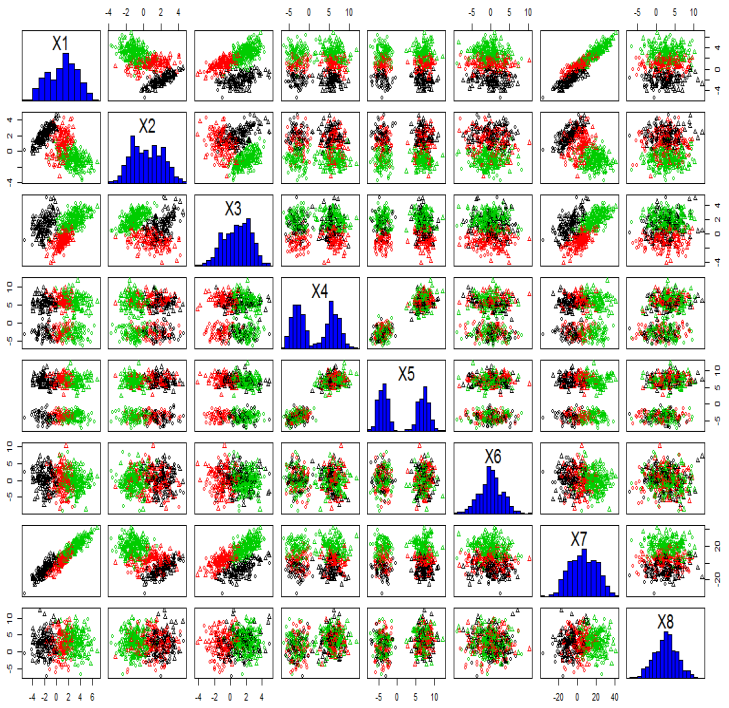


Figure 7.8: Multivariate simulated data, the combination of colours (red, black or blue) and symbols (circle or triangle) divided the data into six subgroups.

When we applied our model-based clustering (where the EM-algorithm was initialized using k-means algorithm) to all the variables (without any variable selection), our method identified five clusters as optimal clusters with 84.2% correct classification; whereas R&D method ("MCLUST") identified eight optimal clusters with true classification 74.8%. However, when we performed the variable selection (dropping irrelevant variables), our

proposed method identified X_1, X_2, X_3, X_4 and X_5 as clustering relevant, and six corresponding optimal clusters with true classification 98.2%; whereas R&D variable selection method ("clustvarsel") identified only X_4 and X_5 as clustering relevant and two corresponding optimal clusters with classification of 43.4%. The results are reported in table 7.1 on page 168.

Table 7.1: Performance of the proposed method using simulated data

Raftery & Dean method			Our method		
Selected variables	Optimal clusters	Classification	Selected variables	Optimal clusters	Classification
All variables	8	74.8%	All variables	5	84.2%
X_4 & X_5	2	43.4%	X_1, X_2, X_3, X_4 & X_5	6	98.2%

Second simulated data

We simulated a second dataset with different scenario to the first simulated data (figure 7.8), in which only one set of relevant variables and majority of the variables are irrelevant (noisy). The simulated data consists of three clustering relevant (X_1, X_2 and X_6) and six (X_3, X_4, X_5, X_7, X_8 and X_9) noisy clustering irrelevant variables (normally distributed). There is no correlation between the clustering relevant and irrelevant variables. The relevant subset (X_1, X_2 and X_6) divided the entire dataset into two distinctive subgroups. The data is displayed in figure 7.9 on page 169.

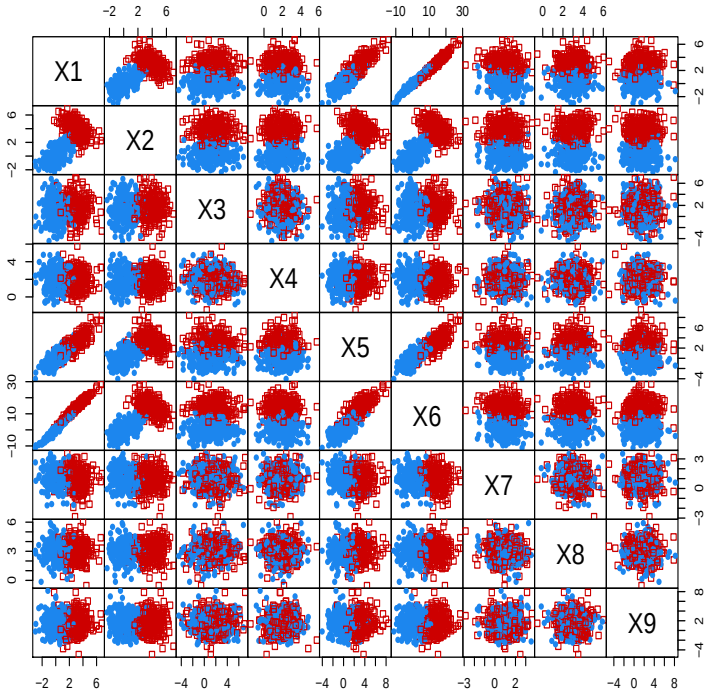


Figure 7.9: Multivariate simulated data, the colours (red and blue) divided the data into two subgroups.

Table 7.2: Performance of the proposed method using simulated data

Raftery & Dean method			Our method		
Selected variables	Optimal clusters	Classification	Selected variables	Optimal clusters	Classification
All variables	3	77.2%	All variables	2	99.8%
X_1, X_2 & X_6	2	99.4%	X_1, X_2 & X_6	2	99.4%

When we applied our model-based clustering (where the EM-algorithm was initialized using k-means algorithm) to all the variables (without any variable selection), our method identified two clusters as optimal clusters with 99.8% correct classification; whereas R&D method ("MCLUST") identified three optimal clusters with true classification 77.2%. However, when we performed the variable selection (dropping irrelevant variables), our proposed method identified X_1, X_2 and X_7 as clustering relevant, and two corresponding optimal clusters with true classification 99.4%; whereas R&D variable selection method ("clustvarsel") also identified X_1, X_2 and X_7 as clustering relevant and two corresponding optimal clusters with classification of 99.4%. The results are reported in table 7.2 on page 169.

Third simulated data

Several random datasets were simulated using "clusterGeneration" R package [214]. This package has a function (genRandomClust) which generates random clusters based on degree of separation index (sepVal), number of clusters, and number of relevant and noisy (irrelevant) variables. The index (sepVal) represents the separation degree between the clusters, which ranges from -0.99 (highly overlapped clusters) to 0.99 (well separated clusters). We generated several datasets (403 observations in each dataset) which have three clusters, five relevant and three noisy variables; however, the degree of separation (sepVal) between clusters varies across these datasets. The datasets are displayed graphically across the clusters in appendix A on page 200 (Figures: a - e). The main aim of this scenario is to assess the performance of the proposed method in identifying the optimal clusters based on the degree of separation of the clusters.

Table 7.3: Performance of the proposed method using simulated data

Raftery & Dean method			**Our method**		
Selected variables	**Optimal clusters**	**Classification**	**Selected variables**	**Optimal clusters**	**Classification**
sepVal=-0.2					
All variables X_1, X_2, X_4 & X_8	No clusters 3	- 80.89%	All variables X_1, X_2 , X_4 & X_8	No clusters 3	- 84.86%
sepVal=0.005					
All variables X_1, X_2 , X_4, X_7 & X_8	3 3	95.53% 95.29%	All variables X_1, X_2 , X_4 & X_8	2 3	71.96% 96.28%
sepVal=0.06					
All variables X_1, X_2, X_4 & X_8	3 3	97.02% 97.02%	All variables X_1, X_2 , X_4 & X_8	3 3	97.52% 97.27%
sepVal=0.1					
All variables X_1, X_2, X_4 & X_8	3 3	99.01% 99.01%	All variables X_1, X_2 , X_4 & X_8	2 3	99.26% 99.01%
sepVal=0.2					
All variables X_1, X_2, X_4, X_7 & X_8	3 2	99.75% 99.50%	All variables X_1, X_2 , X_4 & X_8	3 3	99.50% 100.00%

When we applied the proposed method (where the EM-algorithm was initialized using k-means algorithm) to all the variables (without any variable selection) and to the selected

variables, the proposed method appeared to perform well in identify the optimal clusters as the degree of separation between the clusters increases; and R&D method also showed similar pattern. However, where the degree of separation (sepVal) was 0.005, using all the variables, our method (the EM-algorithm was initialized using k-means) identified two clusters rather than three with 71.96% true classification (although R&D identified the optimal clusters correctly with 95.53% true classification); but when the EM-algorithm was initialized using kmedoids, our method identified three clusters and the true classification improves to 96.03%. This shows that how the initialization of the EM-algorithm affects the identification of the optimal clusters in model-based clustering. The results from R&D and the proposed methods are displayed in table 7.3 on page 170 across the degree of separation (sepVal).

7.8.3 Example 2: Seeds data

To compare the proposed method performance on the real dataset, we used a well know seeds real dataset from "UCI Machine Learning Repository", which commonly researchers use for assessing the performance of their new methods.

Seeds dataset comprised kernels belonging to three different varieties of wheat, such as Kama, Rosa and Canadian; each type consists of 70 elements (samples). They were randomly selected for the experiment, and displayed in figure 7.10 on page 172. The data is described in detail previously here [148], and the data is publicly available on this website: https://archive.ics.uci.edu/ml/datasets/seeds.

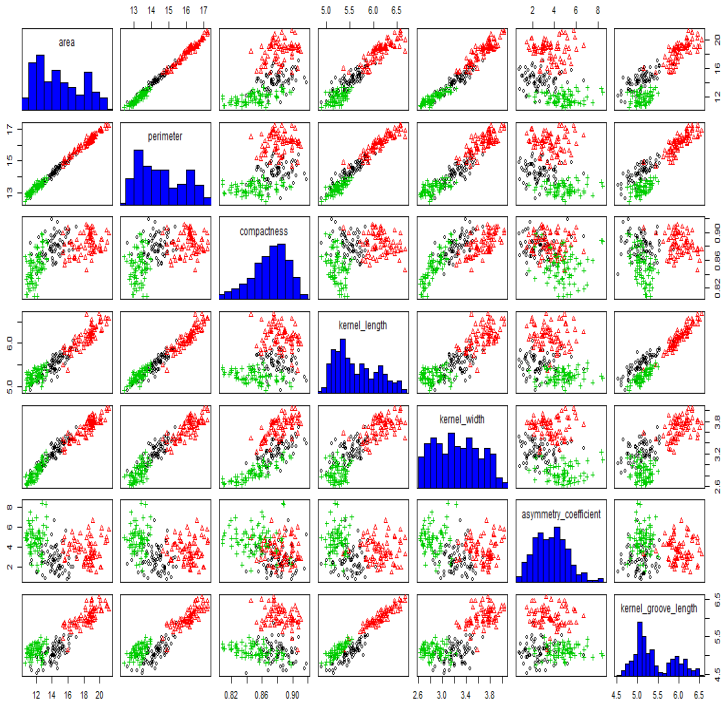


Figure 7.10: Seed dataset, each colour represent three different varieties of wheat (subgroup)

We applied the proposed method to the above seeds real dataset, and compare its performance with R&D method. When we apply our model-based clustering (the EM-algorithm was initialized using k-means algorithm) to all the variables (without any variable selection), our method identified four optimal clusters with true classification 74%; whereas R&D method identified four optimal clusters as well, but with true classification 66%. However, when we performed the variable selection, our method identified $kernel - groove - length$, $perimeter$, $kernel - length$, $area$ & $kernel - width$ and three corresponding optimal clusters with 94.3% true classification; where as R&D variable selection method identified only $perimeter$, $area$ and $compactness$ as clustering relevant and seven corresponding optimal clusters with classification of 47.4%. The results are reported in table 7.4 on page 172.

Table 7.4: Performance of the proposed method using seeds real dataset

Raftery & Dean method			**Our method**		
Selected variables	Optimal clusters	Classification	Selected variables	Optimal clusters	Classification
All variables	4	66%	All variables	4	74%
perimeter, area & compactness	7	47.4%	kernel-groove-length, perimeter, kernel-length, area & kernel-width	3	94.3%

7.8.4 Example 3: Wine data

Furthermore, we used wine real dataset from "UCI Machine Learning Repository" to compare performance of the proposed method with R&D method. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars (the data are depicted in figure 7.11 on page 173). The analysis determined the quantities of 13 constituents found in each of the three types of wines. The data is previously described in detail here [149], and are publicly available here: https://archive.ics.uci.edu/ml/datasets/Wine.

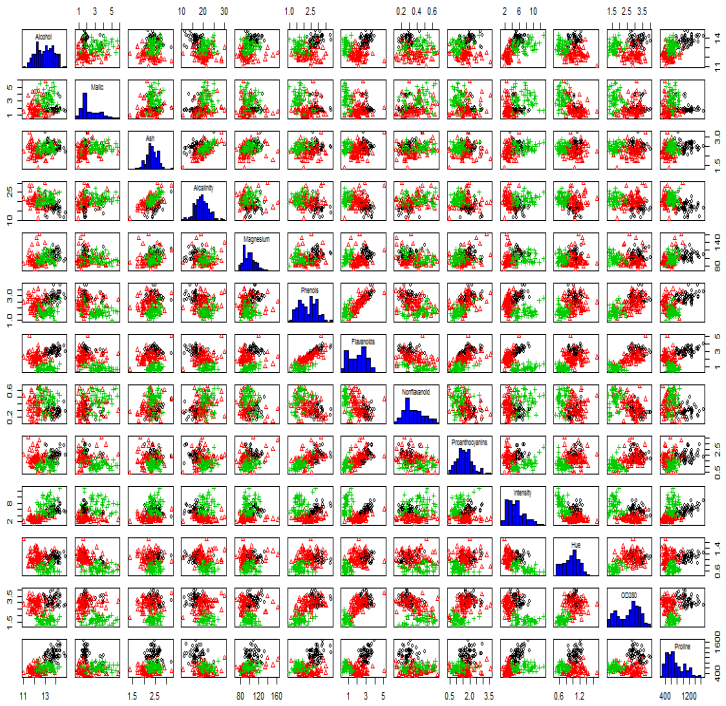


Figure 7.11: Scatterplot matrix of wine data with points marked (coloured) according to the known wine types (subgroup)

We applied the proposed method to the above wine real dataset, and compare again its performance with R&D method. When we apply our model-based clustering (the EM-algorithm was initialized using k-mediods algorithm), to all the variables, our method identified two optimal clusters with 69.1% true classification, whereas R&D identified seven optimal clusters with a classification of 65.7%. However, when we performed the variable selection, our method selects $OD280$, $Flavanoids$, $Phenols$, $Nonflavanoid$, $Proline$, $Magnesium$, $Intensity$, Hue, $Malic\ Proanthocyanins$ as clustering relevant variables, and three corresponding optimal clusters with classification of 95.5%. Whereas R&D variable selection method identified $Malic$, $Proline$, $Flavanoids$, $Intensity$, $Magnesium$,

Alcalinity & *Alcohol* as clustering relevant, and five corresponding optimal clusters with 80.3% true classification. These results are reported in table 7.5 page 174.

Table 7.5: Performace of the proposed method using wine real dataset

Raftery & Dean method			**Our method**		
Selected variables	Optimal clusters	Classification	Selected variables	Optimal clusters	Classification
All variables	7	65.7%	All variables	2	69.1%
Malic, Proline, Flavanoids, Intensity, Magnesium, Alcalinity & Alcohol	5	80.3%	OD280, Flavanoids, Phenols, Nonflavanoid, Proline, Magnesium, Intensity, Hue, Malic & Proanthocyanins	3	95.5%

7.8.5 Example 4: Kim's Simulated data

Kim and his colleague simulated a multivariate data, 300 observations and six variables, and depicted in figure 7.12 on page 174. They did not discuss how many optimal clusters exist in the dataset and left for open discussion. The details are published here [150]. Other studies used these dataset and attempted to identify the clustering relevant and optimal clusters, such as here [151].

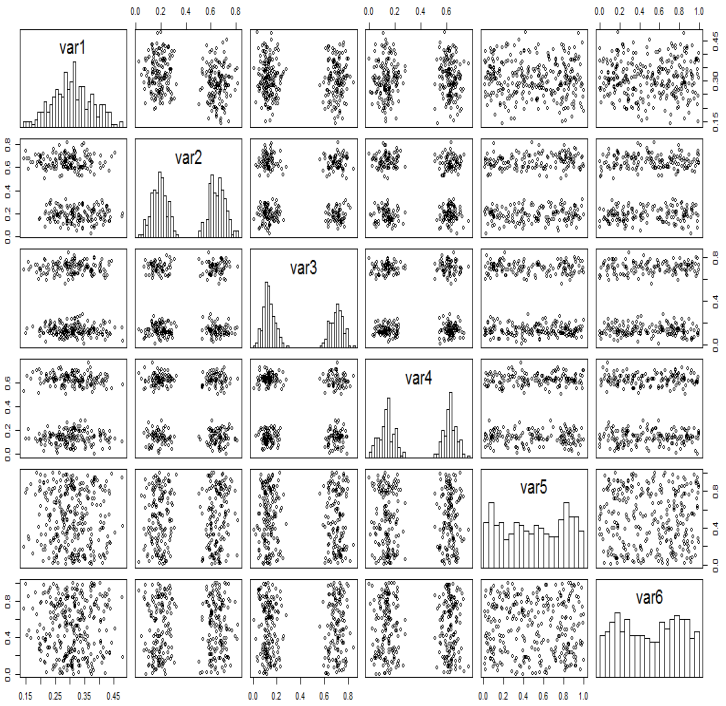


Figure 7.12: Kim's simulated dataset

We applied our method to these dataset and var_2, var_3 and var_4 were selected as clustering relevant variables, and eight corresponding optimal clusters were identified.

R&D method also identified the same relevant variables and optimal number of clusters.

7.8.6 Example 5: Severe Refractory Asthma data

The proposed method is applied to cluster 349 subjects from British Thoracic Society severe refractory asthma registry. Previously a two-stage approach (factor and cluster analyses) was applied to this data, and five optimal clusters were identified [128]. As the data was suffering with missingness, Newby and colleague [128] imputed the missing values. Thereafter, those continuous variables that are missing in less than 30% of the subjects were included in factor analysis, and five factor scores were extracted and used as input into clustering algorithm. The following variables satisfy the criteria and used as input into factor analysis: BMI, hospital admission in the last 12 month, pre FEV_1 predicted, pre FVC% predicted, pre FEV_1/FVC ratio, blood eosinophils, IgE, rescue steroid courses in the last 12 months, Beclomethasone Dipropionate (BDP) equivalent dose inhaled corticosteroids and age at onset of symptoms. For the list of variables used and detailed steps of the multiple imputations, readers should be referred to Newby et al [128].

The factor and cluster analyses approach might not be the best for this data as the variables were not strongly correlated, and are not well represented by the retained five factors. As observed in the table 1, most of the variables' variances (>80%) were not well explained by these retained factors except FEV_1, FVC and FEV_1/FVC. The corresponding factor scores which generate from these factors are also not good representative of all the observed variables (see chapter 3 for factor score formula and how it depends on th factor loadings), therefore the two-stage approach (factor and cluster analyses) might not be the best for this data.

Table 7.6: Varimax rotated factor loadings of sputum mediators at exacerbation

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	C^1	U^2
BMI (kg/m^2)	-0.09	0.13	0.11	0.24	-0.22	0.14	0.86
Admission12m (n)	0.02	0.02	0.39	-0.04	-0.02	0.15	0.85
FEV_1 (%)	0.74	0.65	-0.03	0.05	0.01	0.98	0.02
FVC (%)	0.97	0.02	0.02	-0.03	0.01	0.93	0.07
FEV_1/FVC (%)	0.08	0.96	0.02	0.00	-0.01	0.92	0.08
Blood eosinophils (%)	0.00	-0.06	0.09	0.04	0.31	0.11	0.89
IgE (kU/L)	0.09	-0.03	0.06	-0.25	0.21	0.12	0.88
Rescuest12m (n)	0.01	0.06	0.44	-0.07	0.05	0.21	0.79
BDP equivalent (μg)	-0.06	-0.06	0.26	0.05	-0.18	0.11	0.89
Age onset (years)	-0.05	0.09	-0.13	0.37	0.04	0.17	0.83

C^1 = Proportion of total variation accounted for by the common factors (common variance)
U^2 = Proportion of total variation not accounted by the common factors (unique variance)
Abbreviations: BMI = body mass index; Admission12m=hospital admission in the last 12 month; FEV_1= forced expiratory volume in the first second; FVC= forced vital capacity; Rescue steroids12m= rescue steroid courses in the last 12 months; BDP=Beclomethasone Dipropionate; IgE= immunoglobulin E; 1 (kU/L)= 2.4 (ng/mL)

The proposed method was applied to these data to identify clustering relevant variables and corresponding optimal clusters. Therefore, four variables (IgE, blood eosinophils, BDP equivalent and age at onset of symptoms) were selected as clustering relevant variables, and four corresponding optimal clusters were identified. These clusters have demonstrated clinical utility with cluster-specific pattern. The clinical characteristics across the identified clusters are presented in table 7.7 on page 176.

Table 7.7: Statistical summaries of "severe refractory asthma" clusters which were identified using the new variable selection and clustering method

Variable	Cluster 1 (n=63)	Cluster 2 (n=19)	Cluster 3 (n=129)	Cluster 4 (n=138)	ANOVA P-value
BMI (kg/m^2)	29.45 (0.84)	25.26 (1.24)	29.47 (0.59)	29.37 (0.51)	0.05
FEV_1 (%)	67.17 (3.24)	64.97 (5.54)	65.37 (1.96)	67.05 (2.07)	0.92
FVC (%)	79.81 (2.6)	87.52 (5.3)	82.17 (1.92)	82.35 (1.56)	0.54
FEV_1/FVC (%)	64.87 (1.89)	58.68 (3.47)	61.97 (1.31)	63.43 (1.32)	0.37
Blood eosinophils (%)	0.36 (0.04)	1.07 (0.24)	0.45 (0.04)	0.33 (0.03)	<0.0001
IgE (kU/L)	89.25 (10.4)	1938.55 (403.01)	116.55 (7.75)	359.49 (25.56)	<0.0001
Rescue steroids12m (n)	3.76 (0.48)	5.05 (0.9)	4.89 (0.42)	4.49 (0.36)	0.38
BDP equivalent (μg)	954.76 (15.46)	1646.26 (191.57)	1916.81 (15.89)	2274.02 (111.8)	<0.0001
Age onset (years)	30.15 (2.29)	23.37 (5.34)	29.83 (1.73)	21.38 (1.55)	<0.0001

Abbreviations: BMI = body mass index; Admission12m=hospital admission in the last 12 month; FEV_1= forced expiratory volume in the first second; FVC= forced vital capacity; Rescue steroids12m= rescue steroid courses in the last 12 months; BDP=Beclomethasone Dipropionate; IgE= immunoglobulin E; 1 (kU/L)= 2.4 (ng/mL)

The advantage of applying this approach compare to the previous method is that it identified very few clustering relevant variables, which are easy to measure them in a clinic and would be helpful for subjects selection for clinical trials and assignment of subject to subgroup in which he/she could respond to treatment, and validation of the identified clusters.

7.8.7 Example 6: Asthma and COPD Sputum Cytokines

As the sputum cytokines (which discussed in the first part of this thesis) are strongly correlated and have internal pattern, the two-stage approach (i.e. using factor score as input into clustering algorithm) appeared as best method for this type of data. However, to assess the performance of the proposed method for this kind of data, it was applied to asthma and COPD sputum mediators at exacerbation state to identify the relevant observed cytokines and corresponding optimal clusters. Twenty mediators (such as IL-1β, IL-5, IL-6, IL-6R, IL-8, IL-10, IL-13, CXCL-10, CXCL-11, CCL-2, CCL-3, CCL-4, CCL-5, CCL-13, CCL-17, CCL-26, TNFα, TNF-R1, TNF-R2 and VEGF), were included in the variable selection algorithm. Then TNFα, CXCL-11, CXCL-10, IL-5, IL-10, IL-13, CCL-4 , CCL-5, CCL-13, CCL-26 and TNF-R2 were selected as clustering relevant variables. Thereafter, using these selected variables; two clusters were identified as optimal. These two clusters represented the subjects with elevated Th-2 and proinflammatory profiles, respectively. In which subjects in cluster 1 has elevated level in these Th-2 mediators (such as CCL-27), whereas cluster 2 subjects have elevated level in these proinflammatory mediators (such as IL-1β, TNFα), and the data are displayed in table 7.8 on page 178. However, the subgroup which represents the Th-1 mediators high is not revealed using the new method although it was captured using the two-stage technique (factor and cluster analysis) in the previous analysis (see chapter 6). Thus, for variables which have strong correlation (internal patterns) still the two-stage method might be the best.

Table 7.8: Statistical summaries of asthma and COPD biological clusters which were identified using the new variable selection and clustering method

Variable	Cluster 1 (Asthma=13; COPD=39)	Cluster 2 (Asthma=18; COPD=34)	P-value
IL-1β	36.5 (23.9 - 55.9)	920.8 (587.1 - 1444)	<0.0001
IL-5	1.9 (1.2 - 2.9)	1.5 (1 - 2.2)	0.49
IL-6	185.1 (111.5 - 307.4)	717.7 (488.7 - 1054)	<0.0001
IL-6R	170.7 (125.5 - 232.1)	883.3 (636.7 - 1225.4)	<0.0001
IL-8	2282.3 (1707.8 - 3050)	10192.1 (8065.9 - 12878.6)	<0.0001
IL-10	1.9 (1.8 - 2.1)	24.9 (16.2 - 38.1)	<0.0001
IL-13	11.7 (10 - 13.7)	10.4 (9.1 - 11.8)	0.25
CXCL-10	394.2 (260.5 - 596.7)	773.1 (413 - 1447.2)	0.08
CXCL-11	27.9 (16 - 48.7)	37.2 (15 - 92.2)	0.6
CCL-2	350.4 (276.9 - 443.5)	614.2 (414.7 - 909.7)	0.018
CCL-3	37.9 (27.8 - 51.6)	120.1 (81 - 178.1)	<0.0001
CCL-4	758.3 (541 - 1062.8)	1473.3 (955.2 - 2272.3)	0.02
CCL-5	4 (3.2 - 5.1)	24.9 (18.2 - 34.2)	<0.0001
CCL-13	20.6 (16.7 - 25.6)	16.3 (12.7 - 20.8)	0.15
CCL-17	21.8 (14.3 - 33.4)	6.5 (4.3 - 9.6)	<0.0001
CCL-26	7.1 (4.8 - 10.4)	4.5 (3.3 - 6.1)	0.08
TNFα	3 (2 - 4.6)	104.8 (70 - 156.8)	<0.0001
TNF-R1	620 (485.6 - 791.5)	3344.7 (2589.9 - 4319.4)	<0.0001
TNF-R2	247 (181.3 - 336.6)	1730.4 (1348.8 - 2219.9)	<0.0001
VEGF	1070.6 (947.2 - 1210)	1885.8 (1571.8 - 2262.6)	<0.0001

7.9 Discussion

In this study, a new variable selection method for the model-based clustering was proposed to generalize the R&D variable selection approach (with the hope that variable selection or dropping irrelevant variables in clustering may improve the observations classification into the correct subgroups). The new method searches for latent (structure) among the entire variables and divides the variables into independent subsets on the bases of their internal structures (correlations). In addition, it accounts the linear and non-linear relationship between the clustering relevant and irrelevant variables.

The model-based clustering method (which is used in this study) is optimized using EM-algorithm. The EM-algorithm is initialized using several heuristic clustering algorithms such as k-means, k-medoids or fuzzy k-means. In addition, unconstrained variance-covariance matrix is used and the singularity issues (not positive definite) in the variance-covariance matrix is addressed using an iterative approach. The positive definite of the variance-covariance matrix is tested using Choleski decomposition [152]. If the matrix is not positive definite, the algorithm keeps adding 1% to the diagonal of the matrix till it becomes a positive definite.

The proposed approach appears as promising in its ability to unearth the clustering

relevant variables and corresponding optimal clusters, and starts to outperform the R&D method, which is very encouraging. The only limitation we observed thus far is that it is unable to reveal optimal clusters for dataset which have strong correlation (for this type of data even R&D did not perform well), and our method is also computationally infeasible especially with the increase of the number of observations and variables. However, its speed can be improved massively by using a parallel programming approach (i.e. dividing the task into available processors and execute the sequences of instructions in parallel across the clusters or/and variables' subgroups). In addition, presently this study is focusing on the model-based clustering, but in future it will be extended to heuristic clustering techniques (with some modifications) in which it might account for dataset in which the number of variables is greater than the number of observations, and variables which have strong correlation such as cytokines and gene expression.

In summary, this study showed that including irrelevant variables in clustering could add unnecessary noise and hide existing optimal subgroups. In addition, it also showed that removing variables without properly checking (for globally and locally relevant variables), could cause losing potential clusters as we observed in the R&D method. Although our method appears to choose more variables, it outperformed the R&D method in identifying the relevant variables and optimal number of clusters in the simulations and real data shown.

Chapter 8

Thesis Conclusion and Future Direction

8.1 Objectives

This thesis has two main parts. In the first part, robust statistical techniques were applied to model the biological heterogeneity of asthma and COPD jointly, and subsequently three stable and three exacerbation clusters were identified with different proportions of overlap between the two diseases. In the second part, a new method for variable selection in model based clustering was developed, and was applied to simulated and real dataset. In this chapter, the overall findings, contributions and limitations of the thesis are summarized, and the future direction of the project is also outlined.

8.2 Part One

The first part of the thesis is an application in which the biological heterogeneity of asthma and COPD was modeled jointly using robust statistical techniques. Subsequently, three biological subgroups, separately at stable and exacerbation states with different proportions of overlap between the two diseases were identified. The identified subgroups have clinical and biological implications that may contribute to the personalized medicine, patient selection for clinical trials and adjustment of the standard care treatments targeting each subpopulation.

8.2.1 Stable Biological Subgroups of Asthma and COPD

At stable state, distinct and overlapping biological (sputum mediators) subgroups of asthma and COPD were identified using two-stage statistical technique (factor and cluster analyses). This showed that asthma and COPD have clear distinctive demographic and

lung-function characteristics, but they have considerable common biological features at stable state. Patients from both diseases were categorized into three clinically and biologically relevant subgroups using clustering algorithm. These subgroups were asthma dominant (Cluster 1), asthma and COPD overlap group (Cluster 2), COPD dominant group (Cluster 3); in which cluster 1 was characterized by increased T_H2 inflammatory mediators (such as IL-5, IL-13 and CCL-26) with eosinophilic inflammation. Cluster 2 was the overlap subgroup with increased neutrophilic inflammation, high proportion of bacterial colonization and elevated proinflammatory mediators (such as IL-1β and TNFα). Whereas, cluster 3 was COPD-predominant group with mixed eosinophilic and neutrophilic inflammation and elevated proinflammatory cytokine levels (such as IL-6).

Although cluster 1 was characterized by increased level of T_H2 mediators and eosinophils cellular profiles, but the potential connection between these mediators and eosinophils was not fully understood in which whether the T_H2 cytokines mediate/cause the eosinophilic inflammation. In previous study, Stirling et al (2001) suggested that IL-5 enhances asthma by increasing number of eosinophils [153], but its contribution to COPD is unknown. Cluster 2 (overlap) was also characterized by high level in T_H1 derived and proinflammatory cytokines (such as IL-1β, IL-6R, IL-8, IL-10, CCL-3, CCL-4, CCL-5, CXCL-10, TNFα, and VEGF) and high proportion of bacterial colonization with neutrophilic inflammation. In this cluster these cytokines may have mediation/direct effect in causing the neutrophilic inflammation as there was a significant positive correlation between neutrophilic and IL-1β only in this subgroup. However, these pattern need to be replicated in bigger study, and experimented in a laboratory. This study is an observational from single center, and did not test for causality or mediation effect using mathematical approach due to the lack of enough data within the cluster. In the previous study it has been suggested that IL-1β may enhance asthma and COPD [154], but the mechanism was not fully explained. This group is also characterized by high level of bacteria colonisation compared to the other two clusters. Previous studies suggested that infections (viral and bacterial) may contribute to the pathogenesis and progression of COPD [36], and was reported an association between bacterial colonization and airway inflammation [39], and between airway bacteria load and decline in FEV_1 in COPD patients [40]. Whereas in asthmatic studies, it has been suggested that bacterial organisms may increase airway hyperresponsiveness and inflammation [41], and asthmatics with neutrophilic inflammation are commonly culture-positive for Haemophilus influenza [42, 43], which may suggest the potential role of bacteria presence (especially H. influenza) in the lower airway in the

continuation of neutrophilic airway inflammation [44]. The third cluster is characterized by high levels in IL-6 and CCL-2 mediators, and majority of the subjects were COPD patients. They tend to be older, male dominant with shorter duration of disease, and low lung-function measurements and high pack-year history. The mechanisms of IL-6 and CCL-2 in these airway diseases were not clearly understood, but it has been reported that IL-6 may increase inflammation in asthma and COPD [155]. Whether this subgroups could benefit from anti-IL-6 drug is unknown and require further big observational study and mechanistic experiment in a research laboratory.

In conclusion, this study showed that separation between asthma and COPD at stable state is clearly visible, but there is also an overlap between the diseases at biological level. Asthmatics who are high in T_H2 derived cytokines with eosinophilic inflammation and COPD subjects who have elevated level of proinflammatory cytokines with low neutrophilic cell counts as two distinct airways diseases, but the two diseases overlap when both have elevated level in T_H1 derived and proinflammatory cytokines with neutrophilic inflammation and high proportion of bacterial colonization. Although the underlying potential mechanism is not completely understood, this study may bring new definitions that acknowledge the overlap and highlight the similarities and differences between the two diseases when they are stable. In addition, it may bring attention to the potential contributions of cytokines in classification asthma and COPD phenotypes, which might yield new insights that could benefit future efforts in these airway diseases research, diagnosis, prevention using more personalized intervention approach.

8.2.2 Validation subgroups

The stable subgroups were validated on an independent asthma and COPD study using two approaches (such as discriminant analysis, and combination of disease status and IL-1β cut off), and identified three subgroups with similar patterns of clinical characteristics and mediator profiles across the subgroups as the test clusters, which suggested the stability of the clusters in an independent subjects. Although there are slight differences in the proportion of asthma and COPD between the test and validation subgroups, there are no significant differences in the proportions across the groups. In addition, although the two classifier techniques performed well in validating the study, they need to be replicated in bigger studies before using as a standard approach for subjects' assignment into relevant subgroups.

8.2.3 Exacerbation Biological Subgroups of Asthma and COPD

Similarly, a two-stage technique (cluster and factor analysis) was also applied to model the biological heterogeneity of asthma and COPD using the sputum mediators at exacerbation state. Subsequently, three biological clusters were identified. The first two clusters (cluster 1 and 2) represented the overlap of asthma and COPD, and the third cluster (cluster 3) was COPD predominant subgroup; in which cluster 1 was with eosinophilic inflammation and increased T_H2 inflammatory mediators; cluster 2 was with elevated T_H1 cytokines and increased proportion of subjects with Firmicutes and Streptococcus (phylum Firmicutes); and cluster 3 was clinically chronic bronchitis with neutrophilic inflammation and increased proportion of bacterial colonization and Proteobacteria, and elevated level of proinflammatory mediators.

8.2.4 Similarities between the Stable and Exacerbation Subgroups

The patterns of clinical and biological characteristics observed across the stable and exacerbation subgroups are very consistent; in which subjects who have elevated level of eosinophils have increased level of T_H2 derived cytokines. The subgroups that have high neutrophils cell-counts also have elevated level of proinflammatory mediators and increased proportion of bacterial colonization. However, asthma and COPD subjects appeared to have more similarity on the biological mediators at exacerbation compared to stable state.

This study showed that T_H2 profile is rather associated with eosinophilic inflammation than with neutrophilic or bacterial colonization. On the other hand, neutrophilic inflammation (non-eosinophilic) was characterized by a cytokine profile featuring raised in proinflammatory such as IL-1β and TNFα and increased proportion of bacterial colonization. This pattern highlights that the non-eosinophilic phenotype including neutrophilic represents a major part of asthmatic and COPD population, which may suggests that the non-eosinophilic inflammation is characterized by different molecular mechanisms than eosinophilic inflammation. For example, these T_H2 cytokines were not increased in non-eosinophilic asthma/COPD that may indicate there are other non-T_H2 cytokines (such as T_H1 and proinflammatory) play a role in causing/mediating these non-eosinophilic inflammations (such as neutrophilic inflammation).

Although the reasons why there is such a variation in neutrophilic and eosinophilic inflammations in asthmatic and COPD patients remain unclear but may be linked to the

level of the cytokines profiles and bacterial infection. However, the mechanistic of this pattern requires to be experimented in a laboratory.

8.2.5 Limitations

The specific study limitations have been discussed within each chapter. However, the general limitation of the applied work is that only subjects with severe asthma and COPD who attended a secondary care setting (from a single center) were included, and thus might not be representative of a more generalized population. We acknowledge that our findings cannot be extrapolated to mild and/or moderate asthma or mild COPD but are confident that our populations are representative of our broader secondary care respiratory patient population. Further studies are required to include healthy controls, larger populations including those with mild disease from multi-centres for generalization. In addition, the exacerbations were moderate and findings cannot be extrapolated to severe exacerbations. Furthermore, the description of an exacerbation in this study was relied on several parameters; in which a patient has to experience a change in symptoms and recognise that if is different from baseline (stable) symptoms, and has to report to a medical practitioner. Thereafter, a medical practitioner will assess that whether the reported symptoms are different to the patient's baseline symptoms and requires medical therapy with corticosteroids and antibiotics. However, the medical practitioner does not know whether these symptoms relate to an exacerbation or whether the treatment will work as there is no simple biomarker available to identify exacerbation from stable state. Although, this definition is consistent with the current literature but it has some shortfalls which may have an impact on the findings. In general, the basis of an exacerbation relies on subjective reporting by patients and subjective assessments made by physicians, and the current definition of exacerbation also does not account the underlying cause of the exacerbation, the treatment response or the complex psychological or social influences that exist in patients with chronic respiratory disease. However, identifying a standard biomarker threshold that could predict an exacerbation from stable state may aid in the solving some of these complexity.

Although the longitudinal stability of the clusters was not tested directly in this study, the biological heterogeneity was modeled separately at stable and exacerbation states and revealed similar patterns of the clinical and biological characteristics across the identified subgroups at both states. The patterns show the relationship (pathological connection) between the mediators and other clinical characteristics especially with cellular profiles is

not time and state (stable or exacerbation) dependent.

Furthermore, in this study only mediators which are detectable in more than 50% of the population were included, and may have some influence on the findings, but these mediators represent quite well the overall profiles (panel) of the cytokines (i.e. Th-2, Th-1 and proinflammatory). So the findings, particularly the connection between elevated cellular profiles (eosinophilic and neutrophilic) and Th-2 high or Th-2 low (i.e. high in Th-1 or proinflammatory) is interesting observations, and deserves further investigation in other population in order to develop a new cytokine-based therapeutic targeting each subpopulation similar to anti-IL-5. Anti-IL-5 appeared to be effective in reducing the eosinophilic inflammation in these subjects who have high level of T_H2 cytokines and eosinophilic cell-counts [58, 72, 73]. However, in future a large number of mediators can be assessed using similar statistical approach if new technology is manufactured to measure these undetectable mediators.

Thus far, more than 45 papers cited our published paper (which was generated from this thesis [1]) and Dirkje and colleague [152] discussed the paper and pointed out that some of the demographic and treatments may have effect on the results. However, the effects of these suspected confounding variables were assessed empirically on the factor scores (which were used for constructing the clusters) and found no evidence. Although this is an observational study, the findings were highly unlikely influenced/confounded by these possible asthma and COPD confounding characteristics such as standard treatments and smoking status. This study just revealed novel biological clusters which do not rely on clinical and demographic characteristics in which the mechanisms are not yet understood.

Finally, in this study the patterns were revealed using mathematical algorithms, but not explore the mechanisms underlying this strong relationship (especially the connection between mediators and cellular profiles in both stable and exacerbation states). However, future large longitudinal observational studies from multi-centers and/or experimental studies are required for better understanding of the potential mechanisms of the observed patterns in each subpopulation. In addition, although the clusters at stable were validated on an independent study using two techniques, the exacerbation clusters are required to be validated in similar format to assess the stability of the clusters in an independent population.

8.2.6 Summary of Clinical Findings

In this study the biological subgroups of asthma and COPD were identified using statistical techniques, and revealed further information that was hidden at disease level (across asthma versus COPD group comparison) in the clinical characteristics, biological mediators and microbiome community. These observations may have clinical utilities for targeted medicine and/or may help to understand better the potential mechanism and heterogeneity of the diseases, and encourage further mechanistic studies. For example, it has been observed that subjects who have elevated level in eosinophil inflammation have increased level in T_H2 derived mediators (cytokines). In addition, the subjects who have elevated level in neutrophils cell-counts have increased level in proinflammatory and bacterial colonization at both stable and exacerbation, and increased proportion of Proteobacteria at exacerbation. Furthermore, at exacerbation subjects who have elevated level in T_H1 cytokines have increased proportion of Firmicutes at phylum level and Streptococcus (phylum Firmicutes) at genus level. These observations are novel and the pathological connections/interactions between the characteristics are not fully understood yet in which whether the increase in inflammatory mediators causes/mediated the cellular airways inflammations. Therefore, further experimental studies are required to understand the relationship between the inflammatory mediators (cytokines), cellular profiles, microbiome community and other characteristics in each subpopulation.

In conclusion, there are limited data in the literature to show the patterns of a range of inflammatory mediators and cellular profiles across asthmatics and COPD at both stable and exacerbation states. So our study is the biggest to show the patterns of broad spectrum of characteristics across asthma and COPD at both stable and exacerbation which may extend our knowledge in the field by showing the relationship between clinical, mediator and microbiome in each subgroup beyond the comparison at disease level. This study explored the biological heterogeneity in asthma and COPD, and identified subgroups which may calls for targeted cytokine-based treatments or/and mechanistic study for selected sub-population. In general, this study may bring attention to the potential contributions of cytokines in classification of asthma and COPD phenotypes, which might yield new insights that could benefit future efforts into diagnosis, prevention, patient selection for clinical trials and development of personalized intervention (treatment-specific anti- inflammatory therapies).

8.2.7 Statistical Methods to Model the Biological Heterogeneity of Asthma and COPD

In this study, the performances of several statistical methods were compared on simulation study to identify the right approach to model the biological heterogeneity (using sputum cytokines) of asthma and COPD jointly. Modeling the heterogeneity of the diseases using the sputum mediators (cytokines) was not straightforward as these mediators have strong correlations (internal patterns), and based on these patterns they were partitioned into several independent subgroups. Thus, for this type of data, using factor scores (derived from factor analysis) as input into k-means clustering algorithm outperformed the alternative approaches in the simulation study. Thereafter, this approach was applied to model the biological heterogeneity of asthma and COPD independently at stable and exacerbations states. Subsequently, three stable and three exacerbation independent clusters with different proportions of overlap between the two diseases were identified.

This approach (using factor scores as input into clustering algorithm) works very well in a situation where there are strong correlations between the observed variables as it accounts for the internal patterns (underlining structures) of the variables in partitioning the observations into distinctive subgroups. This means that the correlations of the observed variables are accounted within each cluster, and able to identify subgroup with elevated level of variables that have similar pathways (strong correlation). Whereas, in the standard clustering techniques (i.e. using the observed variables as input into the clustering algorithms), the internal patterns (correlation among the variables within a cluster) are usually ignored, which assumes a local independence (uncorrelated) between the variables within each cluster.

However, prior to applying this technique (using factor scores as input into clustering algorithm) to cluster observations, the internal patterns of the observed variables should be assessed using graphical visualization. If there is evidence of observed variables' subgroups (internal patterns), factor analysis (with varimax rotation) should be implemented to these variables to extract the factor scores for each observation. In addition, it is better to be investigated the variables communalities (the shared variance explained by the retained factors) before extracting the factor scores. If most of the variables have communalities less than 50%, it would be better to use the standard approach (i.e. using all the measured/observed variables as input into the clustering algorithm) instead of using factor scores. In such situation, using the factor scores as input into clustering algorithm

may under-represent the information exists in the entire variables, and its consequence could be quite serious. However, if the variables that have communalities less than 50% are few (e.g. less than 25% of the entire variables) then these variables could be removed from the factor analysis prior to extracting the factor scores as these variables may hide the optimal clusters exist in the dataset.

In addition, if the observed variables have strong correlation, the retained factors (based on the screeplot or eigenvalue above one) may capture most of the information exist in the observed variables. Thus, using the corresponding factor scores (which are retained using the above criteria) as input into a clustering algorithm would be robust (although this is a heuristic approach and hard to justify mathematically). However, it could be a dangerous practice to use this approach in a situation where there are weak correlations between the variables, which may lead to the extraction of very few factors that greatly under-represent the entire information exists in the observed variables as observed in these studies [91, 128].

Furthermore, in a situation where the factor scores are used as input into clustering algorithm, it would be better to predict the identified clusters using linear discriminant analysis using the observed variables. This approach is very helpful in identifying variables that have significant contribution in discriminating the clusters, and may aid for future validation or subject's assignment to the identified subgroups. In addition, this may give a general overview that how well the clusters are distinctive with respect to the observed variables as usually the identified clusters utilities are interpreted based on the patterns of the observed variables instead of on the patterns of the factor scores (latent variables).

Moreover, K-means clustering algorithm was used for identifying subjects' subgroups in this study. It's a popular non-probabilistic data-clustering algorithm which is not prone to model overfit (when the number of parameters becomes greater than the observations), and is in the top three algorithms been used in the last 10 years for similar clustering purpose. However, one of its main drawbacks is, despite the various suggestions (indexes), there is no commonly acceptable methodology on how to compare models with different numbers of clusters in order to identify the optimal number of clusters. The "Elbow" (Scree plot) technique was used in this analysis to choose the possible number of clusters by incorporating the suggestion from variables subgroups and clinical and biological plausibility of the clusters. The rationale behind the Elbow methods is that the clusters before the break (inflation) in the graph (which displays the within cluster variation against the number of clusters) capture most of the information, and the clusters after the break is

very near to some of the existing ones and do not add any further information. There is a probabilistic model-based approach, which is analogous to "K-means clusters on factor scores" for identifying clusters with underlying structure, which is known as Factor Mixture Model (FMM). FMM uses latent class analysis for classification and factor analysis for identifying the underlying structure, here the optimal number of clusters is chosen based on Akaike information criterion (AIC), Bayesian Information Criterion (BIC) and other measures of goodness fit. This approach has an advantage over the method applied in this analysis as all subjects who have at least one cytokines can be included in the analysis, and the clusters can be adjust for other possible confounding covariates (continuous and categorical). However, this approach is prone to overfit, and computationally intensive and currently could handle maximum up to three factors, and only is implemented in Mplus (a commercial statistical software).

8.2.8 Future Direction

The asthmatic and COPD subjects who participated in this study have been followed up and further measurements such as gene expression and CT (x-ray computed tomography) scan of their small and large airways will be available soon. Therefore, the patterns of such characteristics will be assessed across asthma and COPD, and across the identified biological subgroups at both stable and exacerbation state, and further subgroups will be explored using statistical algorithms. In addition, patterns of the microbiome communities at phylum and genus levels at stable state are in the process of being extracting from the existing samples, and will be assessed across stable and exacerbation, and further microbiome subgroups of the diseases will be investigated using robust statistics techniques.

In addition, the approach that uses factor scores (derived from factor analysis with varimax rotation) as input into a clustering algorithm appeared to outperform the alternative approaches, particularly in a situation where there are strong correlations between the observed variables. Therefore, this approach will be developed into an algorithm by incorporating visualization techniques (using graphical techniques for assessing the internal structure of the variables), and a criterion for removing noise variables (e.g. variables which have high error-terms compared to their communalities in factor analysis) prior to extracting factor scores for clustering. Thereafter, the identified clusters will be predicted using discriminant analysis to identify variables that have significant contribution in discriminating the subgroups, The may aid for future subject assignment to the relevant group using few observed variables. In addition, the issue of missing data will be handled.

The algorithm will be written in R (statistical software platform) and will be publicly available as an open source.

In this study the mediators provided valuable information regarding the clinical subgroups during stable and exacerbations states, and indicate clinically important tools that may result in better understanding of the disease process and lead to superior management strategies and direct personalized therapy. However, the patterns were revealed using mathematical algorithms, but not explore the mechanisms underlying the relationships observed among the characteristics in each subgroup (especially the connection between cytokines and cellular profiles at both stable and exacerbation states). Although the longitudinal work and validation develop some understandings in the stability of the subgroups, the overall mechanisms is very complex, and it is not known how these cytokines alter expression of multiple inflammations in these airway diseases. In other chronic inflammatory diseases, such as rheumatoid arthritis and inflammatory bowel diseases, blocking some of these cytokines has proven to be of clinical benefit, so there has been considerable interest in determining whether the same approach might also be useful in inflammatory airway diseases and these can be experimented in research laboratory. Therefore, further mechanistic work to assess the connection between the biological clusters and cellular inflammation and microbiological are required. For example, this approach may provide a basis to investigate whether exacerbations due to airway inflammations and bacterial infection could be prevented by treating with cytokine-based drugs (such as anti-Th-2) and antibiotic, respectively.

8.2.9 Benefit and Limitation of Cluster Analysis in Medical Research

Cluster analysis is an unsupervised multivariate technique that aims to classify a sample of subjects on the basis of a set of variables into a number of distinctive (discrete) subgroups. The observations within each cluster are more similar (homogeneous) to each other than between clusters (minimizing within group variation and maximizing between groups variation). There is a major advantage of using cluster analysis to characterize heterogeneity population in medical research. For example, it may provide new insights and enhance our understanding of the diseases complexity (where the characterisation of patients on the basis of clusters of symptoms can be useful in the identification of an appropriate personalised/stratified approach of therapy). In addition, it aids to direct future mechanistic studies, structure detection and generating hypothesis (exploring the deeper relationships between variables in each subpopulation). Furthermore, it may identify a novel potential biomarker that could predict treatment response and could be also useful for patients' selection in clinical trial (to maximize the chances of success) and adjusting the standard treatments targeting each subgroup.

Although cluster analysis has advantage in understanding complex dataset, it has several limitations in which researchers and readers should be aware. The robustness of the identified clusters could be influenced due to several things: such as because of the structure of the variables (as most of the clustering algorithms assume multivariate normality), outliers, missing values (usually the standard techniques assume only complete case analysis), choice of measure of similarity, choice of the input variables, and uncontrolled confounding factors. Thus, researchers should consider carefully and be cautious whether a specific clustering algorithm is a suitable and meaningful approach for the dataset at hand before drawn any conclusions from the result. Many of these algorithms are greedy (i.e. the optimal local solution is always taken in the hope of finding an optimal global solution in which sometimes quite impossible especially when the clusters are not well separated). Applying an appropriate clustering algorithm based on the type of the input data is the best approach; for example, some clustering approach allows detection of irregular clusters (i.e. those which have poorly defined shapes), and some are sensitive to outliers compared to others.

Confounding Variables in Cluster Analysis

Some potential confounding factors (such as treatments and demographic characteristics) may influence cluster membership, and should be accounted whilst clustering the subjects. For example, if the potential confounding factors (which are not included in the clustering algorithm) affect these variables that are used as input into cluster analysis, then the robustness (validity) of the identified clusters could be questionable and the analysis may lead to distorted results which may have a profound consequence. Most of the common clustering algorithms, particularly these heuristic approach (such as k-means and hierarchical) do not have capability to adjust for confounding factors. However, the effect of these suspected confounding factors can be assessed empirically. For example, confounding factors can be investigated as follows: by assessing whether the variables used as input into the clustering algorithm are associated with the confounding factors and/or the confounding factors are not the reverse effect of these input variables. If no evidence is found, then it can be concluded that the identified clusters are highly unlikely could be influenced/confounded by these possible confounding factors. However, we have to be aware that there should be unknown factors which could influence the variables (which used as input into clustering algorithm) and the results should be interpreted cautiously till the observed relationships are justified in an experimental study.

Missing Data Issues in Cluster Analysis

Missing data is the main issue in any statistical analysis including cluster analysis due to the fact that considerable number of subjects are more likely to be excluded from the analysis because of the missingness. The main cause of missing values in medical research are not fully understood. However, they could be due to the fact that the questionnaires were not filled properly, invalid values were recorded, undetectable measurements (e.g. a patient may not produce enough sputum in which it may affect the extraction of all the possible sputum mediators and microbiome communities) and so on.

Most of the common clustering techniques do not handle missing data and require only complete cases (subjects without missing values). In cluster analysis, excluding substantial number of subjects with missing values may underpowered the robustness of the results as the subjects which included in the analysis may not represent these excluded. There are several formal and informal approaches have been implemented to address the issue of missing data in clustering. Such as replacing the missing values by the average of the

available data (mean imputation); but the issue with this approach is that a large numbers of missing values are more likely to be similar to each other as the same values are assigned for these missing values. In addition, if other approaches such as stochastic models are used for imputation, randomization are added to the data in which subjects with missing data makes them highly unlikely to cluster (group) together. In general, applying imputation techniques in cluster analysis to impute missing values might be dangerous as this approach may introduce unrealistic (artificial) (dis)similarity in the dataset.

Dealing with missing data in cluster analysis is at its early stage, and requires an attention from researchers in which, at least, these informal approaches aforementioned should to be tested rigorously using simulation studies whether they are robust under the assumptions of Missing At Random (MAR), Missing Completely At Random (MCAR) and Missing Not at Random (MNAR) mechanisms. Although the assumptions could be difficult to check, the simultaneous investigation of the missing data patterns and the observed values may allow a better understanding of the missing data mechanisms. In addition, the percentage of the imputed values under certain missingness mechanisms should be quantified in order to identify robust clusters using the imputed data. However, in situation where the imputation is impossible, at least the representation of these subjects which were excluded from the analysis should to be assessed (whether they have similar characteristics to these subjects were included) for the generalization of the results.

Selection of Input Variable in Cluster Analysis

The choice of the input variables in cluster analysis is quite controversial, especially when the choice is not justified mathematically and/or clinically. For example, the identified clusters could be very dependent on the variables which are included in the clustering algorithm. Although it seems the more information that exists for individuals would be better for clustering, adding non-informative (irrelevant) clustering variables may not make a basic change in the identification of the optimal clusters except hiding the existing subgroups. Therefore, including these non-informative variables as input into the clustering algorithm could be a harmful (which may add unnecessary noise and hide the optimal clusters) as the clustering non-informative variables may dominate the effect of these clustering informative variables.

The general objective of variable selection in clustering is to maximize the identification of the optimal number of clusters using minimum number of clustering relevant variables. Particularly in medical research it would be useful for subject selection for clinical trials,

and adjustment of the standard care treatments targeting each subgroup based on the state (threshold) of these few relevant variables.

However, identifying an appropriate method to choose the relevant variables for clustering (without losing additional information with these excluded) is an ongoing study. The common approaches of choosing input variables for clustering algorithm are: based on the clinical utilities of the variables by eliminating these redundant ones (which is very subjective), and reducing the dimension of the observed variables using factor or principal component analyses (PCA) and use the corresponding scores as input into clustering algorithm, or using the highest-loading observed variables (after factor analysis), which may not be robust approaches for dataset which have no strong correlations (see chapter 3 for details).

Selection of Optimal Number of Clusters

In cluster analysis, it is necessary to select the optimal number of clusters although it is not always straightforward. There are several informal (subjective) approaches especially in heuristics clustering techniques, and more formal (using BIC) in model-based clustering.

Most of the existing clustering techniques rely on heuristic methods that are based on similarity or dissimilarity distance measures (such as k-means and hierarchical clustering algorithms). This approach is computationally feasible and available in most open-source and commercial statistical software and not prone to model overfitting. However, one of its main drawbacks is that, despite the various suggestions (indexes), there is no commonly acceptable methodology on how to compare models with different numbers of clusters in order to identify the optimal number of clusters. In addition, the choice of the optimal clusters is very subjective, and usually depends on the clinical interpretation and plausibility of the clusters.

An alternative method is a model-based clustering, in which a more formal statistical procedure can be implemented to choose the optimal clusters (mixture components) using likelihood approach such as BIC. However, although the optimal clusters could be chosen based on the BIC but not always guaranteed for their clinical interpretation. Even the EM-algorithm (which uses to optimise the model-based clustering) is prone to initialization where it could converge to a local maxima rather than the global maximum, in which there may be multiple optimal solutions (clusters) for a single dataset, especially when the clusters are not well separated. Thus, the choice of the optimal clusters is a "trade off" between the criterion and interpretablility of the clusters.

Stability of Clusters

Testing stability of clusters in independent studies, and overtime within the same study is very important before drawn a conclusion based on a single cross-sectional findings. The stability of a clustering solution can be validated on an independent studies using supervised techniques (such as linear discriminant analysis). However, assessing how the clusters (e.g. identified at baseline) would be stable overtime is not straightforward although it can be investigated heuristically using supervised techniques (by predicting the baseline clusters using the follow-up data and check for the misclassification), but this approach does not answer the cause (mechanism) of that misclassification as it does not account the trajectories of the variables overtime (temporal changes). An alternative methods are needed to identify the follow-up clusters (which accounts the within and between subjects variations) by adjusting the baseline clusters and identify the impact of each variable in keeping or misclassifying the subjects. If this approach is far from exist, as it has been implemented in this study, clustering can be implemented separately at stable and exacerbation states and compare the patterns (relationship between variables in each subgroups). For example, in this study, we found that the relationship between mediators and cellular profiles in each subgroup are similar at both stable and exacerbation states, which shows that these patterns are time and state (stable or exacerbation) independent (invariant).

Conclusion

Application of cluster analysis in medical research may provide a valuable information, and identify clinically important subgroups which lead to the superior management strategies and more personalized therapy. There are a wide range of robust clustering techniques which work well, but researchers should be aware of their caveats and known issues associated with them in order to identify a more robust clusters. As each clustering technique behaves differently for different type of data, users should have to check the assumptions of each clustering algorithm in order to identify the right approach for the dataset at hand. In addition, prior to clustering, a rigour investigation (screening) of the variables should be implemented using appropriate techniques (such as using graphical visualization and data reduction) in order to get a guidance which approach to use. Furthermore, it is better to implement post clustering analysis (compare the patterns and relationship of variables within and between the clusters), and investigating the stability of the clusters in an independent study and overtime.

Cluster analysis is generally an exploratory approach which helps to discover hidden patterns in very complex dataset. However, the current approaches have several limitations; so a new clustering technique is needed which is not prone to model overfitting, and has a flexibility in handling missing data, selecting input variables, adjusting for potential confounding factor and testing stability of the clusters overtime.

8.3 Part Two

8.3.1 Proposed Variable Selection for Model-based Clustering

In this part, a new method for variable selection in model based clustering is proposed. This method extends the technique that was originally developed by R&D [129]. The proposed algorithm searches for latent structures in the entire variables, and splits the variables into two or more independent subsets if there is evidence of internal patterns among the variables. Thereafter, forward variable selection algorithm is performed in each variables' subset, and all the relevant variables are aggregated from each subset and corresponding global optimal clusters were identified. Furthermore, each global cluster is treated as independent dataset and further search for local clustering relevant variables from these globally clustering irrelevant variables are performed. If local clustering relevant variable is identified, that variable is included with these globally relevant variables. Finally, using all the aggregated relevant variables the corresponding optimal clusters (which may exist in the entire dataset) are identified.

Overall, this approach relaxes the unrealistic assumptions of the relationship between clustering relevant and irrelevant variables (where R&D method assumes), in a situation where these variables are uncorrelated (independent). In addition, it accounts for the non-linear relationship between clustering relevant and irrelevant variables by breaking down the entire dataset into possible subgroups.

In this proposed method, an independent model-based clustering is developed in which it is optimized using EM-algorithm. The EM-algorithm is initialized using several heuristic clustering algorithms such as k-means, k-medoids or fuzzy k-means. In addition, the singularity issue of the variance-covariance matrix is addressed by replacing the diagonal of the matrix with small positive value (i.e. 1%). The performance of this approach was assessed in the simulation and real dataset and compared with R&D method.

8.3.2 Limitation

The proposed method appeared to outperform the R&D method, but it requires complete cases at the moment in which it couldn't impute missing values although it is possible to impute the missing values and use the imputed variables as input into the algorithm, similar to what Newby and colleague [128] have applied for the ATS asthmatic data. In addition, the method performed better in data which do not have strong correlations among the variables as the sputum cytokines. However, for data with strong correlation among the variables such as sputum cytokines still the two stage (factor and cluster analyses) would be the best in identifying the clusters with underlying variable profiles. In addition, the new method is prone to overfit if the number of the parameters is greater than number of observation as it is a model-based clustering technique.

8.3.3 Conclusion

The proposed method outperformed the R&D method and showed that clustering using clustering relevant variables improved substantially the identification of optimal clusters exist in the dataset. In addition, although no variable selection is performed (i.e. using all the observed variables as input into clustering), the proposed clustering method appeared to outperform the model-based clustering technique that was used in the R&D method. The possible explanations for the limitations in the R&D method could be because of the initialization of the EM-algorithm (which is initialized using hierarchical clustering algorithm) or/and the unnecessary constrains of the variance-covariance matrices in their model-based clustering technique in order to avoid the singularity issues.

8.3.4 Future Direction

In future, this work will be expanded to account the situation where the variables are greater than the observations, without being prone to overfitting. In addition, the algorithm will be extended to heuristic clustering algorithm and handle missing data issues (includes missing data imputation algorithm and implement simultaneously with the variable selection and clustering). In addition, at the moment the proposed method perhaps does not perform adequately in uncovering the optimal clusters in the dataset which have strong correlations between the variables (such as sputum mediators); so the algorithm will be extended to account for these type of data.

Appendix A

Figures of Simulated Data

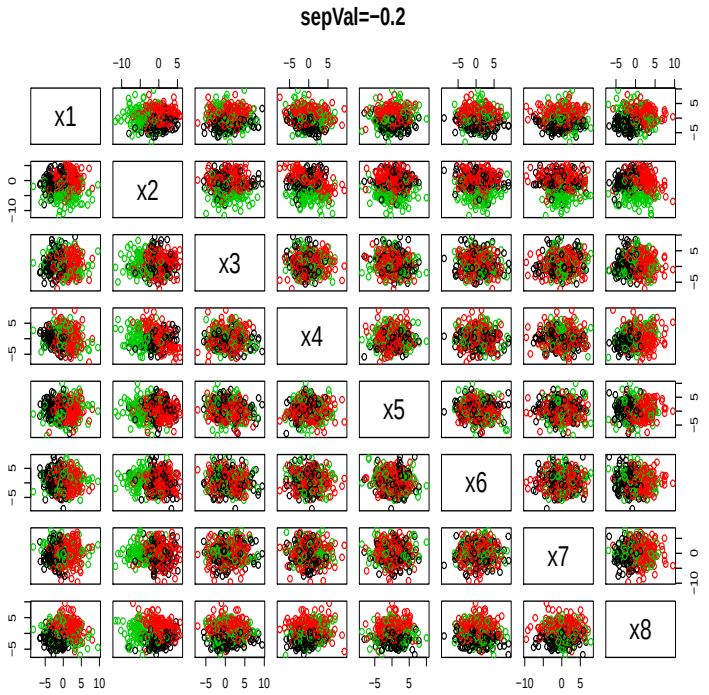


Figure (a): multivariate simulated data where the degree of separation (sepVal) between the clusters is -0.2, the colours represent the three clusters.

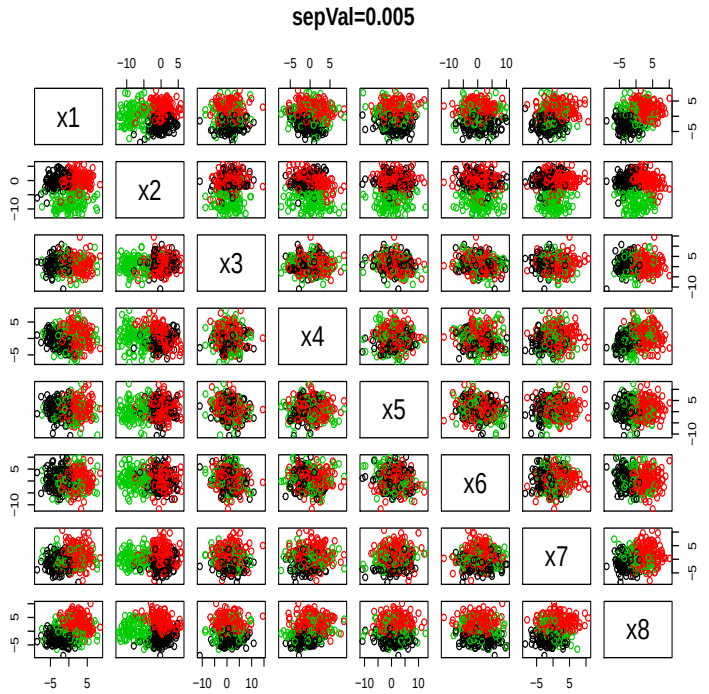


Figure (b): multivariate simulated data where the degree of separation (sepVal) between the clusters is 0.005, the colours represent the three clusters.

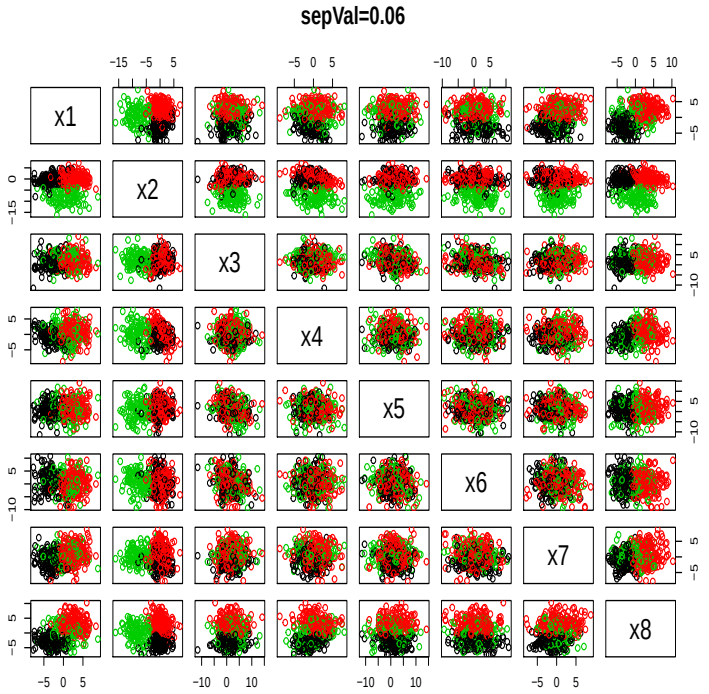


Figure (c): multivariate simulated data where the degree of separation (sepVal) between the clusters is 0.06, the colours represent the three clusters.

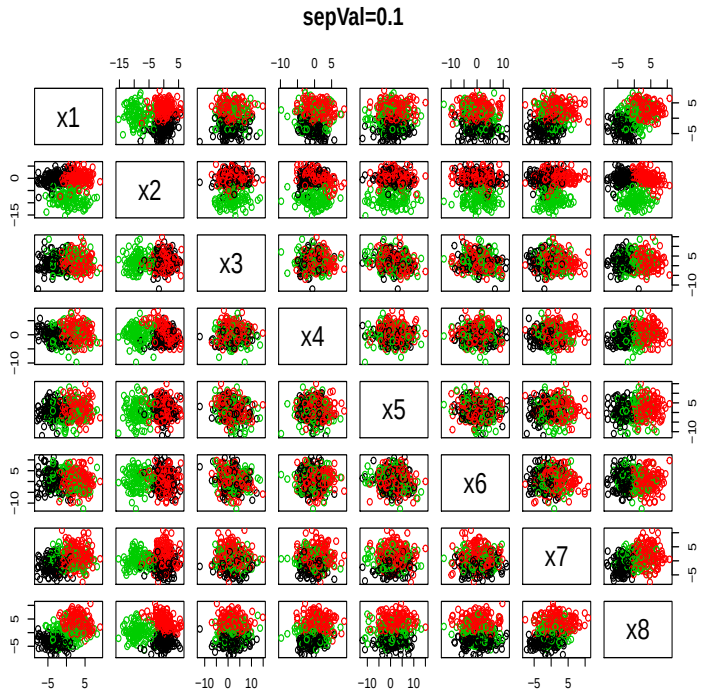


Figure (d): multivariate simulated data where the degree of separation (sepVal) between the clusters is 0.01, the colours represent the three clusters.

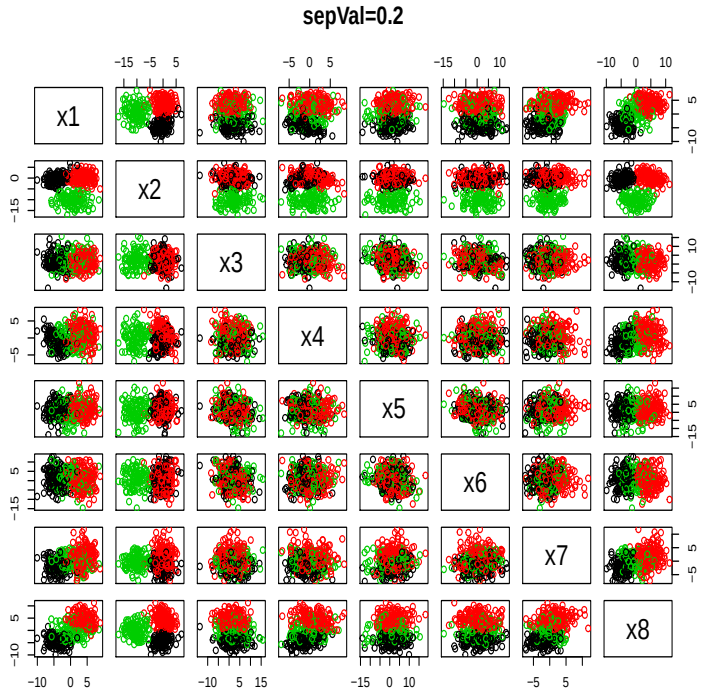


Figure (e): multivariate simulated data where the degree of separation (sepVal) between the clusters is 0.2, the colours represent the three clusters.

Appendix B

R-code for the Proposed Variable Selection Method

In this session, all the R codes for the proposed variable selection method for model-based clustering are presented.

```
##########################################################
# Fix singularity of variance-covariance  matrix in    #
# model-based clustering which fits using EM-algorithm #
##########################################################

sigmaFixer <- function(sigma, ...){
  sigma <- as.matrix(sigma)
  r <- dim(sigma)[1]
  fixrate <- 0.01
  covfixmat <- array(1,c(r,r)) + fixrate*diag(1,r)
  min_limit <- .Machine$double.eps*10

  if (!all(is.finite(sigma))){
    warning("covariance matrix is not finite")
  }

# Enforces the squareness and symmetricity
  nsigma <- sigma - ((sigma - t(sigma))/2)
  iter <- 0

 # Checking the covariance matrix is not positive definite
  while (postDef(nsigma) == 0 & (iter < 10000)){
```

```
23
24      iter <- iter + 1
25      d <- diag(nsigma)
26      if (any(d <= min_limit)){
27        m <- max(abs(d))*fixrate
28        neg <- min(d)
29
30        if (neg < 0){
31          addit <- (m - neg)*diag(1,r)
32        }
33        else {
34          if (m < min_limit){
35            m <- min_limit
36          }
37          addit <- m*diag(1,r)
38        }
39      nsigma <- nsigma + addit
40      }
41      else {
42      # Increase the diagonal values by 1%
43        nsigma <- nsigma*covfixmat
44      }
45  }
46      # return(list(nsigma, iter))
47      return(nsigma)
48  }
50
51  #testing for postive definite
52  postDef <- function(Sigma){
53    Sigma <- as.matrix(Sigma)
54    q <- try(chol(Sigma)[2], TRUE)
55    if (q==0){
56      t <- 1
57    }
```

```
58    else {
59      t <- 0
60    }
61     return(t)
62  }
67
68  ##############################################################
69  # R code for Gaussian Mixture Model using EM-algorithm
70  ##############################################################
71  #'  gmmEM
72  #'
73  #' This is a model-based clustering method which is
       optimazed using EM algorithm. The EM-algorithm is
       initialized using k-means, k-medoids or fuzzy k-means
       (cmeans) algorithms
74  #' @param x a numeric  matrix or dataframe  for clusterng; c
       is  number of clusters;
75  #' @return  Estimated means and variance-covariance matrix
       in each cluster, and BIC, loglikelihood, number of
       iteration, clusters (class) and number of parameters
76  #' @examples
77  #' gmmEM(iris[,1:4], c=2,initialize = c("kmeans"))
78  #'
79  ##############################################################
80
81  gmmEM <- function(x, c=1:10, initialize = c("kmeans",
       "kmedoids","fuzzykmeans"),iter = NULL, tol = NULL, ...) {
82
83    require(clusterSim)  # for initialization the kmeans
       clustering algorithm
84    require(FactMixtAnalysis) #  for calculating confusion
       matrix
85    require(mvtnorm) # for multivariate density function
86    require(e1071)   # for fussy kmeans
```

```
87
88    x <- as.matrix(x)
89
90    if (is.null(x)) return(NULL)
91
92    #X is non-missing
93    if (any(is.na(x))) {
94      warning("NA's in the dataframe")
95      return(NULL)
96    }
97
98    n <- nrow(x)
99    d <- ncol(x)
100
101   if (c==1 && d==1){
102     j <- 1   # number of clusters
103
104     # initialise the algorithm using k-means, k-medoids or
      fuzzykmeans algorithms
105
106     if (initialize=="kmedoids"){
107       k <- pam(x, c, do.swap=F)
108     }
109     else if (initialize=="fuzzykmeans"){
110       k <- cmeans(x, c, iter.max = 100)
111     }
112       else {
113         k <- kmeans(x, c, nstart = 1)
114       }
115
116     mu <- mean(x[k$cluster == 1])
117     sigma <- sd(x[k$cluster == 1])
118
119     p <- sum(k$cluster == 1)/length(x)
```

```
120
121      loglik <- sum(log(p*(dnorm(x,mu,sigma))))
122
123      p <- (2*j-2 +3*j*d +j*d^2)/2 # parameters or degree of
     freedom ; j = number of clusters
124      bic <- -2*loglik+ p*log(n)
125
126      return(list(mean=mu, sigma=sigma, loglikhood=loglik,
     n=n, BIC = bic, df = p, clusters=j))
127
128    }
129
130    else if (c>=2 & d==1){
131
132  # initialize the EM-algorithm using kmeans, kmedoids or
     fuzzykmeans algorithms
133      if (initialize=="kmedoids"){
134        k <- pam(x, c, do.swap=F)
135      }
136        else if (initialize=="fuzzykmeans"){
137          k <- cmeans(x, c, iter.max = 100)
138        }
139
140      else {
141        k <- kmeans(x, c, nstart = 25)
142      }
143
144      j <- c
145      mu <- lapply (1:c, function(i) mean(x[k$cluster==i,]))
146      mu <- lapply(mu, function(x){replace(x, is.na(x),
     .Machine$double.eps)})
147
148      sigma <- lapply (1:c, function(i) sd(x[k$cluster==i,]))
```

```
149     sigma <- lapply(sigma, function(x){replace(x, is.na(x),
    .Machine$double.eps)})
151
152 #   sigma <- lapply(sigma,function(x)
    replace(sigma,x==0,.Machine$double.eps)) # replacing
    zeros sigma
153
154     p <- lapply (1:c, function(i)  sum(k$cluster==i)/nrow(x))
155     p <- lapply(p, function(x){replace(x, is.na(x),
    .Machine$double.eps)})
156
157     Ntau <- lapply(1:c, function(i)
    p[[i]]*(dnorm(x,mu[[i]],sigma[[i]])))
158     Ntau <- sapply(Ntau, cbind)
159     SUMtau <- apply(Ntau,1,sum)
160     loglik   <- sum(log(SUMtau))
161
162     tol = 1e-06
163     iter <- 0
164     tau <- 0
165
166     # for loop
167     for (i in 1:1000) {
168
169       iter <- iter + 1
170
171       Ntau <- lapply(1:c, function(i)
    p[[i]]*(dnorm(x,mu[[i]],sigma[[i]])))
172       Ntau <- sapply(Ntau, cbind)
173       SUMtau <- apply(Ntau,1,sum)
174
175       tau <- Ntau/SUMtau
176       p <- lapply(1:c, function(i) sum(tau[,i])/nrow(x))
```

```
177       p <- lapply(p, function(x){replace(x, is.na(x),
    .Machine$double.eps)})
178
179       mu <- lapply(1:c, function(i)
    sum(tau[,i]*x)/sum(tau[,i]))
180       mu <- lapply(mu, function(x){replace(x, is.na(x),
    .Machine$double.eps)})
181
182       sigma <- lapply(1:c, function(i) sqrt(sum(tau[,i]
183               * (x - mu[[i]])^2)/sum(tau[,i])))
184       sigma <- lapply(sigma, function(x){replace(x,
    is.na(x), .Machine$double.eps)})
185
186       loglik_0 <- loglik
187
188       Ntau <- lapply(1:c, function(i)
    p[[i]]*(dnorm(x,mu[[i]],sigma[[i]])))
189       Ntau <- sapply(Ntau, cbind)
190       SUMtau <- apply(Ntau,1,sum)
191       loglik   <- sum(log(SUMtau))
192
193       del_loglik <- loglik - loglik_0
194
195       if ((abs(del_loglik) < tol) ||
    (is.nan(abs(del_loglik)))) {
196
197         break
198
199       }
200     }
202
203     prob <- as.data.frame(tau)
204     class <- apply(prob, 1, which.max)
205
```

```
206     par <- (2*j-2 +3*j*d +j*d^2)/2 # parameters or degree of
    freedom; j is number of clusters
207
208     bic <- -2*loglik+ par*log(n)
209
210     return(list(lambda = p, mu = mu, sigma =sigma,
211                 loglikhood = loglik, n=n ,BIC= bic,df=par,
    number_of_iteration = iter,
212                 clusters=j, class = class))
213   }
214
215   else if  (c >= 2 & d >= 2)  {
216
217     #  initialize the algorithm using k-means, fuzzykmeans
    or kmedoids
218     if (initialize=="kmedoids"){
219       k <- pam(x, c, do.swap=F)
220     }
221
222     else if (initialize=="fuzzykmeans"){
223       k <- cmeans(x, c, iter.max = 100)
224     }
225       else {
226         k <- kmeans(x,x[initial.Centers(x, c),])
227       }
228
229     j <- length(unique(k$cluster))
230
231     mu <- lapply (1:c, function(i) apply(x[k$cluster==i,],2,
    mean))
232     sigma <- lapply (1:c, function(i) cov(x[k$cluster==i,]))
233     sigma <- lapply(sigma,function(x) sigmaFixer(x)) #
    fixing singularity
234
```

```
235     p <- lapply (1:c, function(i)  sum(k$cluster==i)/nrow(x))
236
237     tol <- 1e-6
238     iter <- 0
239     tau <- 0
240
241     Ntau <- lapply(1:c, function(i)
    p[[i]]*(dmvnorm(x,mu[[i]],sigma[[i]])))
242     Ntau <- sapply(Ntau, cbind)
243     SUMtau <- apply(Ntau,1,sum)
244
245     loglik   <- sum(log(SUMtau))
246
247     # for loop
248     for (i in 1:1000){
249
250       iter <- iter + 1
251
252       Ntau <- lapply(1:c, function(i)
    p[[i]]*(dmvnorm(x,mu[[i]],sigma[[i]])))
253       Ntau <- sapply(Ntau, cbind)
254       SUMtau <- apply(Ntau,1,sum)
255
256       tau <- Ntau/SUMtau
257       p <- lapply(1:c, function(i) sum(tau[,i])/nrow(x))
258
259       mu <- lapply(1:c, function(i) apply (x, 2, function(x)
    sum(tau[,i]*x)/sum(tau[,i])))
260
261       x_m <- lapply(1:c, function(i) sweep(x, 2,
    mu[[i]],"-"))
262
263       sigma <- lapply(1:c, function(i)
    (t(as.matrix(x_m[[i]]))%*%(as.matrix(x_m[[i]])
```

```
264                       * tau[,i]))/sum(tau[,i]))
265       sigma <- lapply(sigma,function(x) sigmaFixer(x)) #
    fixing singularity
266
267       loglik_0 <- loglik
268       Ntau <- lapply(1:c, function(i)
    p[[i]]*(dmvnorm(x,mu[[i]],sigma[[i]])))
269       Ntau <- sapply(Ntau, cbind)
270       SUMtau <- apply(Ntau,1,sum)
271       loglik   <- sum(log(SUMtau))
272
273       del_loglik <- loglik -  loglik_0
275
276       if ((abs(del_loglik) < tol) ||
    (is.nan(abs(del_loglik)))) {
277
278         break
279
280       }
281
282     }
283
284     prob <- as.data.frame(tau)
285     class <- apply(prob, 1, which.max)
286     par <- (2*j-2 +3*j*d +j*d^2)/2 # parameters or degree of
    freedom
287
288     bic <- -2*loglik+ par*log(n)
289
290     return (list(lambda=p, mu=mu,
    sigma=sigma,loglikhood=loglik,
291                  number_of_iteration=iter, n=n, BIC = bic,
    df=par, class=class))
292   }
```

```
}

#################################################################
#R-code for Linear Regression BIC
#################################################################
#' REGbic
#'
#' This calculates regression BIC
#' @param y is dependent variable as numeric vector, and x
   is independent variable/s  as numeric matrix
#' @return  regression BIC (numeric)
#' @export
#' @examples
#' REGbic()
#################################################################

REGbic <- function (y, x) {

  y <- as.vector(y)
  x <- as.matrix(x)

  if (any(is.na(y))|| any(is.na(x))) {
    warning("NA's in the y or x")
    return(NULL)
  }

  if (any(is.null(y))|| any(is.null(x))) {
    return(NULL)
  }

  p <- ncol(x) + 2

  n  <- length(y)
  fit <- lm(y~x)
```

```
328   sigma <- (sum((summary(fit)$resid)^2)/n)^0.5
329
330   if (ncol(x)==1){
331     REG.bic <- - (-n*log(2*pi)-2*n*log(sigma)-n-log(n)*3)
332   } else {
333
334     REG.bic <- - (-n*log(2*pi)-2*n*log(sigma)-n-log(n)*p)
335   }
336   return(REG.bic)
337 }
340
341 ##################################################################
342 # R-code for Variable Selection in model-based Clustering
      using
343 # Forward  Stepwise Algorithm
344 ##################################################################
345 #'  EMvSel
346 #'
347 #' This selects clustering relevant variables for
      model-based clustering using greedy forward selection
      algorithm
348 #' @param X is a numeric matrix/dataframe, bic is a criteria
      to select a variable as clustering relevant, default is 0
349 #' @return  clustering relevant variables (character)
350 #' @export
351 #' @examples
352 #' EMvSel()
353 ##################################################################
354 # forward feature selection algorithm
355 # Forward greedy search
356 ##################################################################
357
358 EMvSel<- function(X, bic=0) {
359
```

```
360   X <- as.data.frame(X)
361
362   if (is.null(X)) return(NULL)
363
364   #X is non-missig
365   if (any(is.na(X))) {
366     warning("NA's in the dataframe")
367     return(NULL)
368   }
369
370   n <- nrow(X)
371   d <- ncol(X)
372
373   if (is.null(X)) return(NULL)
374
375   #First Step - selecting single variable
376
377   BIC_opt <- rep(NA,d)  # optimal BIC
378   BIC_diff <- rep(NA,d)
379   BIC_one <- rep(NA,d)
381
382 # identify the optimal univariate clusters
383
384    univ.BICs <-lapply(2:3, function(i) try(apply(X,2,
      function(X) gmmEM(X,c=i,initialize =
      c("kmeans"))$BIC),TRUE))
385
386    univ.BIC <- data.frame(t(sapply(univ.BICs,c)))
387
388    univ.BIC <- replace(univ.BIC, is.na(univ.BIC),
      .Machine$double.xmax)
389
390   try(BIC_one <-  apply(X, 2, function(X) gmmEM(X,c=1,
      initialize = c("kmeans"))$BIC), TRUE)
```

```
391
392   BIC_sd <- sweep(-(univ.BIC), 2, FUN = "+", BIC_one) #BIC_d
       <- c(BIC_one - BIC_opt)
393
394   BIC_d <- apply(BIC_sd,2,max) # choosing the highest
       difference
395
396   #Find the variable with the highest BIC difference between
       optimal clusters and no cluster
397
398   v <- max(BIC_d[is.finite(BIC_d)])
399   g <- which(BIC_d==v,arr.ind=TRUE)[1]
400
401   #This is the first selected variable with most univariate
       clustering evidence
402   if (max(BIC_d[is.finite(BIC_d)]) > 0){
403     S <- matrix(c(X[,g]),n,1)
404
405   }
406   else {
407     return(list(VarSel=NULL, var ="No relevant variable"))
408     stop("No relevant variable")
409   }
410
411   # optimal BIC of the first selected relevant variable
412   BIC_S <- min(univ.BIC[g])
413   colnames(S) <- colnames(X)[g]
415
416   if (ncol(X)==1){
417
418     if (BIC_d[g] >= 10){
419
420       return(list(VarSel=colnames(X)))
421     }
```

```
422     if (BIC_d[g] < 10) {
423       #          return(NULL)
424       return(list(VarSel=NULL, var ="No relevant variable"))
425       stop("No relevant variable")
426     }
427   }
429
430   if  (ncol(X) > 1) {
431
432     # N is the matrix of currently irrelevant variables
433     N <- as.matrix(X[,-g])
434     colnames(N) <- colnames(X)[-g]
435
436     #subset is a matrix records the proposed variable,
      optimal BIC of S and difference in BIC for clustering
      versus no clustering on S.
437
438     subset <- matrix(c(colnames(S),round(BIC_S, 4),
      round(BIC_d[g],4),"Yes"),1,4)
440
441     #Second Step - selecting second variable
442     BIC_reg <- rep(0,ncol(N))
443     BIC_joint <- rep(0,ncol(N))
444     BIC_sum <- rep(0,ncol(N))
445     BIC_j <- rep(0,ncol(N))
447
448     #Bivariate joint clustering
449
450     biv.BICs <-lapply(2:6, function(i) try(apply(N,2,
      function(N) gmmEM(cbind(S,N),c=i, initialize =
      c("kmeans"))$BIC),TRUE))
451     err <- sapply(biv.BICs, is, class2="try-error")
452     nulls <- sapply(biv.BICs, is, class2="NULL")
453     biv.BIC_s <- biv.BICs[err==FALSE & nulls==FALSE]
```

```
454     biv.BIC <- data.frame(t(sapply(biv.BIC_s,c)))
455
456     biv.BIC <- replace(biv.BIC, is.na(biv.BIC),
        .Machine$double.xmax)
458
459      # regressing non-relevant variable on relevant variable
460      try(BIC_reg <- apply(N,2,function(N) REGbic(N, S)),TRUE)
462
463      BIC_sum <- BIC_reg + BIC_S
464
465      BIC.df <- sweep(-(biv.BIC), 2, FUN = "+", BIC_sum)
466
467      BIC_diff <-apply(BIC.df,2,max)
469
470      #Choose the variable with the largest difference
471      v <- max(BIC_diff[is.finite(BIC_diff)])
472      g <- which(BIC_diff ==v,arr.ind=TRUE)[1]
473
474      BIC_opt <- apply(biv.BIC, 2, min)
475
476      #add the second best variable if its BIC difference is
        positive (greater than bic)
477      if(BIC_diff[g] > bic){
478        subset <-
        rbind(subset,c(colnames(N)[g],round(BIC_opt[g],4),
479                round(BIC_diff[g],4),"Yes"))
480        j <- c(colnames(S),colnames(N)[g])
481        S <- cbind(S,N[,g])
482        colnames(S) <- j
483        N <- as.matrix(N[,-g])
484      } else{
485
486        subset <-
        rbind(subset,c(colnames(N)[g],round(BIC_opt[g], 4),
```

```
          round(BIC_diff[g],4),"No"))
487
488       return(list(VarSel=colnames(S), Steps=subset))
489       stop()
490     }
491
492     S <- as.data.frame(S)
493     ss <- names(S)
494     N <- X[ , -which(names(X) %in% ss)]
495     N <- as.data.frame(N)
496     colnames(N) <-names(X)[!names(X) %in% names(S)]
497     iter<-0
499
500     while((ncol(N) !=0) & !is.null(ncol(N)) & (iter <
    ncol(X))) {
501
502       iter <- iter + 1
503
504       BIC_reg <- rep(0,ncol(N))
505       BIC_joint <- rep(0,ncol(N))
506       BIC_diff <- rep(0,ncol(N))
507       BIC_opt <- rep(0, ncol(N))
509
510       # identifying the optimal multivariate clusters
511
512       mv.BICs <- lapply(2:6, function(i) try(gmmEM(S[,j],
    c=i, initialize = c("kmeans"))$BIC,TRUE)) #  optimal BIC
    using the relevant variables
513
514       mv.BIC <- data.frame(t(sapply(mv.BICs,c)))
515       mv.BIC <- replace(mv.BIC, is.na(mv.BIC),
    .Machine$double.xmax)
516
```

```
517       mr <- min(sapply (mv.BIC, cbind)) #   BIC for optimal
      cluster
518       MC <- which(sapply (mv.BIC, cbind)==mr,arr.ind=TRUE)[1]
519       MC <- MC+1  #  optimal clusters
521
522       # BIC of optimal clusters
523       try(BIC_opt <- gmmEM(S[,j],c=MC, initialize =
      c("kmeans"))$BIC,TRUE)
524       try(BIC_joint <- apply(N,2, function(N)
      gmmEM(cbind(S[,j],N),c=MC, initialize =
      c("kmeans"))$BIC),TRUE)
525
526       # regressing non-relevant variable on relevant variable
527       try(BIC_reg <- apply(N,2,function(N) REGbic(N,
      S[,j])),TRUE)
528
529       BIC_sum <- BIC_reg + BIC_opt
530       BIC_diff <- BIC_sum - BIC_joint
531
532       #Choose the variable with the largest BIC difference
533       v <- max(BIC_diff[is.finite(BIC_diff)])
534       g <- which(BIC_diff==v,arr.ind=TRUE)[1]
536
537       if(BIC_diff[g] > bic) {
538
539         #if this difference is positive, add this variable
      to S and update the clustering model's BICs
540
541         subset <-
      rbind(subset,c(colnames(N)[g],round(BIC_opt,4),
542                 round(BIC_diff[g],4), "Yes"))
543         j <- c(colnames(S),colnames(N)[g])
544         S <- as.data.frame(cbind(S,N[,g]))
545         colnames(S) <- j
```

```
546
547         ss <-names(S)
548
549         N <- (X[ , -which(names(X) %in% ss)])
550         N <- as.data.frame(N)
551         colnames(N) <-names(X)[!names(X) %in% names(S)]
552
553       } else{
554
555         subset <-
    rbind(subset,c(colnames(N)[g],round(BIC_opt,4),
    round(BIC_diff[g],4),"No"))
556         break
557       }
558     }
559
560     #Lists the selected variables and the matrix of steps'
    information
561     colnames(subset) <-
    c("ProposedVariable","optimal_BIC","BIC_Difference",
    "Relevant?")
562     return(list(VarSel=colnames(S), Steps=subset))
563   }
564 }
567
568 ############################################################
569 # R-code for Relevant Variable Selection in each Variables'
570 # Subset
571 ############################################################
572
573 #' subEMvSel
574 #'
575 #' This function splits the entire variables into
    independent small subsets, and implement the function
```

```
   "EMvSel" to each subset to select clustering relevant
   variables
#' @param Y is a numeric  matrix/dataframe; BIC  is a
   criteria to select a relevant variable, default is 0
#' @return  aggregated clustering relevant variables from
   each variable subset (character)
#' @export
#' @examples
#' subEMvSel()
#'
############################################################
# Subset of variables are identified using factor analysis
   with varimax rotation
############################################################
# The above function (EMvSel) algorithm is applied in each
# variables' subset to identify the relevant variables
############################################################

#Factor Analysis

subEMvSel <- function (Y, BIC=0) {

  require(psych)
  Y <- as.data.frame(Y)
  d <- ncol(Y)

# identify the possible number of factors for the entire
   variables using factor analysis with varimax rotation.
 fact <- lapply(1:d, function(i) try(fa(Y, i, SMC=F, rotate=
   "varimax")$factors,TRUE))

 f <-  max(suppressWarnings(na.omit(as.numeric(fact))))

  if (f >=2){
```

```
604   # if the variables have more than two blocks implement the
       following codes
605     loading <- abs(fa(Y, f, SMC=F, rotate=
       "varimax")$loadings[,1:f])
606     factors <- as.data.frame(loading)
607
608     eigenvalue <- apply( factors,2, function(x) sum(x^2))
609     kk <- as.data.frame(eigenvalue)
610
611     # extracting factors which have total explained variance
       greater than 1
612     eigen1 <- kk[kk > 1,1]
613
614     k <- length(eigen1) # possible number of factors
615
616     if (k > 1){
617       loading <- fa(Y, k, SMC=F, rotate=
       "varimax")$loadings[,1:k]
618       loading <- abs(as.data.frame(loading))
619
620       # identify subset of variables (creating block for
       variables before clustering)
621       subst <- apply( loading, 1, function(x) sample( c(
       colnames(loading)[ which( x == max(x))]),))
622       s1 <- as.vector(subst)
623       s2 <- cbind(s1, variables=names(subst))
624       #dt <- data.table(s2)
625       dt <- as.data.frame(s2)
626       dtt <- cbind(dt, Factor=as.numeric(dt$s1))
627       dtt <- as.data.frame(dtt)
628       datt <- dtt[,2:3]
629
630       f.subset <- as.character()
```

```
631       q <- length(table(datt$Factor)) # final variables' subsets
632
633       for (i in 1:q) {
634         sub <- datt[datt$Factor==i,]
635         rownames(sub) <- sub[,1]
636         subset <- rownames(sub)[rownames(sub) %in% names(Y)]
637         mm <- EMvSel(Y[,c(subset), drop=F], bic=BIC) # feature selection in each block and collect the relevant variables
638         #s <- names(mm$VarSel)
639         f.subset <- c( f.subset, c(mm$VarSel))
640       }
641
642     }
643     else {
644       f.subset <- EMvSel(Y, bic=BIC)$VarSel
645
646     }
647   }
648
649   else {
650     f.subset <- EMvSel(Y, bic=BIC)$VarSel  # extract these matching with the original variables only to avoid null
651   }
652
653   return (f.subset)
654 }
658
659 ##############################################################
660 # R-code for Variable Selection in Model-based clustering
661 ##############################################################
662
663 #' gmmVarSel
```

```
#'
#' This function implements the function (subEMvSel) in each
   global cluster to search further for local relevant
   variables.
#' Z is a numeric matrix/ dataframe; BIC_diff is the
   criteria to select the relevant variable in each cluster,
   default is 10
#' @return final clustering relevant  variables (character)
#' @export
#' @examples
#' gmmVarSel(iris[,1:4])
#' [1] "Petal.Length" "Sepal.Width"  "Petal.Width"
############################################################

gmmVarSel <- function(Z, BIC_diff=10){

  Z <- as.data.frame(Z)

  s <- suppressWarnings(subEMvSel(Z)) #  relevant variables
   selected using algorithm 1

  if (ncol(as.matrix(Z[,!colnames(Z) %in% s])) >= 1 &
   length(s) >0){

    opt_clusterBIC <- lapply(2:3, function(i)
   try(gmmEM(Z[,s], c=i, initialize =
   c("kmeans"))$BIC,TRUE)) #  optimal BIC using relevant
   variables
    opt_clusterBIC <- sapply (opt_clusterBIC, cbind)
    opt_clusterBIC <- replace(opt_clusterBIC,
   is.na(opt_clusterBIC), .Machine$double.xmax)
    vv <- min(sapply(opt_clusterBIC, cbind)) #   BIC for
   optimal cluster
```

```
687      G <- which(sapply (opt_clusterBIC,
     cbind)==vv,arr.ind=TRUE)[1]
688      G <- G+1    #  optimal cluster using relevant variables
689
690      class <- gmmEM(Z[,s], c=G, initialize =
     c("kmeans"))$class # class assigment from algorithm 1
691      Z.new <- cbind(Z, class)
692      c_var <- suppressWarnings(lapply(1:G, function(i)
     try(subEMvSel(Z[Z.new$class==i,!colnames(Z) %in%
     s],BIC=BIC_diff),TRUE))) #  select relevant in each
     cluster
693
694      c_s <-
     as.vector(unique(do.call(rbind,as.list(sapply(c_var,
     cbind))))) #  relevant variable in each cluster
695      cs <- colnames(Z[,colnames(Z) %in% c_s,drop=F])
696      all_s <- unique(as.vector(c(s, cs))) #  all selected
     relevant variables from algorithm 1 and 2
697
698      return(all_s)
699    }
700
701   else {
702     return(s)
703
704   }
705
706 }
```

Bibliography

[1] Michael A Ghebre, Mona Bafadhel, Dhananjay Desai, Suzanne E Cohen, Paul Newbold, Laura Rapley, Jo Woods, Paul Rugman, Ian D Pavord, Chris Newby, et al. Biological clustering supports both "dutch" and "british" hypotheses of asthma and chronic obstructive pulmonary disease. *Journal of Allergy and Clinical Immunology*, 135(1):63–72, 2015.

[2] Brendan J Carolan and E Rand Sutherland. Clinical phenotypes of chronic obstructive pulmonary disease and asthma: recent advances. *Journal of Allergy and Clinical Immunology*, 131(3):627–634, 2013.

[3] Fadia T Shaya, Du Dongyi, Manabu O Akazawa, Christopher M Blanchette, Jingshu Wang, Douglas W Mapel, Anand Dalal, and Steven M Scharf. Burden of concomitant asthma and copd in a medicaid population. *CHEST Journal*, 134(1):14–19, 2008.

[4] Global initiative for asthma guidelines, 2015. URL `www.ginasthma.org`.

[5] Frank C Sciurba. Physiologic similarities and differences between copd and asthma. *CHEST Journal*, 126(2_suppl_1):117S–124S, 2004.

[6] DW Denning, BR O'driscoll, CM Hogaboam, P Bowyer, and RM Niven. The link between fungi and severe asthma: a summary of the evidence. *European Respiratory Journal*, 27(3):615–626, 2006.

[7] Global initiative for chronic obstructive lung disease, 2015. URL `www.goldcopd.org`.

[8] David M Mannino, Earl S Ford, and Stephen C Redd. Obstructive and restrictive lung disease and markers of inflammation: data from the third national health and nutrition examination. *The American journal of medicine*, 114(9):758–762, 2003.

[9] Peggy M Simon, Richard M Schwartzstein, J Woodrow Weiss, and Vladimir Fencl. Distinguishable types of dyspnea in patients with shortness of breath1-3. *Am Rev Respir Dis*, 142:1009–1014, 1990.

[10] Ian Charlton, Gillian Charlton, Judy Broomfield, and Mark A Mullee. Evaluation of peak flow and symptoms only self management plans for control of asthma in general practice. *Bmj*, 301(6765):1355–1359, 1990.

[11] Johanne Coté, Andre Cartier, Patricia Robichaud, Helene Boutin, Jean-Luc Malo, Michel Rouleau, Andree Fillion, Michelle Lavallée, Monica Krusky, and Louis-Philippe Boulet. Influence on asthma morbidity of asthma education programs based on self-management plans following treatment optimization. *American journal of respiratory and critical care medicine*, 155(5):1509–1514, 1997.

[12] Jose M Ignacio-Garcia and Pedro Gonzalez-Santos. Asthma self-management education program by home monitoring of peak expiratory flow. *American journal of respiratory and critical care medicine*, 151(2):353–359, 1995.

[13] Wendy C Moore, Eugene R Bleecker, Douglas Curran-Everett, Serpil C Erzurum, Bill T Ameredes, Leonard Bacharier, William J Calhoun, Mario Castro, Kian Fan Chung, Melissa P Clark, et al. Characterization of the severe asthma phenotype by the national heart, lung, and blood institute's severe asthma research program. *Journal of Allergy and Clinical Immunology*, 119(2):405–413, 2007.

[14] Wendy C Moore, Deborah A Meyers, Sally E Wenzel, W Gerald Teague, Huashi Li, Xingnan Li, Ralph D'Agostino Jr, Mario Castro, Douglas Curran-Everett, Anne M Fitzpatrick, et al. 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, 2010.

[15] MJA Tasche, JC Van der Wouden, JHJM Uijen, BP Ponsioen, RMD Bernsen, LWA van Suijlekom-Smit, and JC De Jongste. Randomised placebo-controlled trial of inhaled sodium cromoglycate in 1–4-year-old children with moderate asthma. *The Lancet*, 350(9084):1060–1064, 1997.

[16] Mark D Eisner, Nicholas Anthonisen, David Coultas, Nino Kuenzli, Rogelio Perez-Padilla, Dirkje Postma, Isabelle Romieu, Edwin K Silverman, and John R Balmes. An official american thoracic society public policy statement: Novel risk factors and the global burden of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 182(5):693–718, 2010.

[17] JRT Colley, WW Holland, and RT Corkhill. Influence of passive smoking and parental phlegm on pneumonia and bronchitis in early childhood. *The Lancet*, 304 (7888):1031–1034, 1974.

[18] Monica Kraft. Asthma and chronic obstructive pulmonary disease exhibit common origins in any country! *American journal of respiratory and critical care medicine*, 174(3):238–240, 2006.

[19] Peter J Barnes. Mechanisms in copd: differences from asthma. *Chest Journal*, 117 (2_suppl):10S–14S, 2000.

[20] Peter J Barnes. The cytokine network in asthma and chronic obstructive pulmonary disease. *The Journal of clinical investigation*, 118(11):3546, 2008.

[21] Tobias Welte and David A Groneberg. Asthma and copd. *Experimental and Toxicologic Pathology*, 57:35–40, 2006.

[22] Roberto Bianchi, Francesco Gigliotti, Isabella Romagnoli, Barbara Lanini, Carla Castellani, Michela Grazzini, and Giorgio Scano. Chest wall kinematics and breathlessness during pursed-lip breathing in patients with copd. *Chest Journal*, 125(2): 459–465, 2004.

[23] S Kesten and KR Chapman. Physician perceptions and management of copd. *CHEST Journal*, 104(1):254–258, 1993.

[24] PM Calverley. Neuropsychological deficits in chronic obstructive pulmonary disease. *Monaldi archives for chest disease= Archivio Monaldi per le malattie del torace/-Fondazione clinica del lavoro, IRCCS [and] Istituto di clinica tisiologica e malattie apparato respiratorio, Università di Napoli, Secondo ateneo*, 51(1):5, 1996.

[25] B Loveridge, P West, MH Kryger, and NR Anthonisen. Alteration in breathing pattern with progression of chronic obstructive pulmonary disease. *The American review of respiratory disease*, 134(5):930–934, 1986.

[26] Samuel Louie, Amir A Zeki, Michael Schivo, Andrew L Chan, Ken Y Yoneda, Mark Avdalovic, Brian M Morrissey, and Timothy E Albertson. The asthma-chronic obstructive pulmonary disease overlap syndrome: pharmacotherapeutic considerations. 2013.

[27] Amir A Zeki, Michael Schivo, Andrew Chan, Timothy E Albertson, and Samuel Louie. The asthma-copd overlap syndrome: a common clinical problem in the elderly. *Journal of allergy*, 2011, 2011.

[28] So Ri Kim and Yang Keun Rhee. Overlap between asthma and copd: where the two diseases converge. *Allergy, asthma & immunology research*, 2(4):209–214, 2010.

[29] Juan José Soler-Cataluña, Borja Cosío, José Luis Izquierdo, José Luis López-Campos, José M Marín, Ramón Agüero, Adolfo Baloira, Santiago Carrizo, Cristóbal Esteban, Juan B Galdiz, et al. Consensus document on the overlap phenotype copd–asthma in copd. *Archivos de Bronconeumología (English Edition)*, 48(9):331–337, 2012.

[30] Marc Miravitlles, Myriam Calle, and Juan José Soler-Cataluña. Clinical phenotypes of copd: identification, definition and implications for guidelines. *Archivos de Bronconeumología (English Edition)*, 48(3):86–98, 2012.

[31] Philippa Shirtcliffe, Mark Weatherall, Justin Travers, and Richard Beasley. The multiple dimensions of airways disease: targeting treatment to clinical phenotypes. *Current opinion in pulmonary medicine*, 17(2):72–78, 2011.

[32] Peter G Gibson, Vanessa M McDonald, and Guy B Marks. Asthma in older adults. *The lancet*, 376(9743):803–813, 2010.

[33] C Magnus Sköld. Remodeling in asthma and copd–differences and similarities. *The clinical respiratory journal*, 4(s1):20–27, 2010.

[34] PG Gibson and JL Simpson. The overlap syndrome of asthma and copd: what are its features and how important is it? *Thorax*, 64(8):728–735, 2009.

[35] Yvonne J Huang, Eugenia Kim, Michael J Cox, Eoin L Brodie, Ron Brown, Jeanine P Wiener-Kronish, and Susan V Lynch. A persistent and diverse airway microbiota present during chronic obstructive pulmonary disease exacerbations. *OMICS A Journal of Integrative Biology*, 14(1):9–59, 2010.

[36] SK Jindal, AN Aggarwal, K Chaudhry, SK Chhabra, GA D Souza, D Gupta, SK Katiyar, R Kumar, B Shah, and VK Vijayan. A multicentric study on epidemiology of chronic obstructive pulmonary disease and its relationship with tobacco smoking and environmental tobacco smoke exposure. *Indian Journal of Chest Diseases and Allied Sciences*, 48(1):23, 2006.

[37] Paul T King, Martin MacDonald, and Philip G Bardin. Bacteria in copd; their potential role and treatment. *Translational Respiratory Medicine*, 1(1):13, 2013.

[38] Daniel C Chambers, Shaan L Gellatly, Philip Hugenholtz, and Philip M Hansbro. Jtd special edition 'hot topics in copd'—the microbiome in copd. *Journal of thoracic disease*, 6(11):1525–1531, 2014.

[39] Mark D Eisner, John Balmes, Patricia P Katz, Laura Trupin, Edward H Yelin, and Paul D Blanc. Environmental health: A global access science source. *Environmental health: a global access science source*, 4:7, 2005.

[40] Tom MA Wilkinson, Irem S Patel, Mark Wilks, Gavin C Donaldson, and Jadwiga A Wedzicha. Airway bacterial load and fev1 decline in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 167 (8):1090–1095, 2003.

[41] Monica Kraft. The role of bacterial infections in asthma. *Clinics in chest medicine*, 21(2):301–313, 2000.

[42] Lisa G Wood, Jodie L Simpson, Philip M Hansbro, and Peter G Gibson. Potentially pathogenic bacteria cultured from the sputum of stable asthmatics are associated with increased 8-isoprostane and airway neutrophilia. *Free radical research*, 44(2): 146–154, 2010.

[43] Jodie L Simpson, Terry V Grissell, Jeroen Douwes, Rodney J Scott, Michael J Boyle, and Peter G Gibson. Innate immune activation in neutrophilic asthma and bronchiectasis. *Thorax*, 62(3):211–218, 2007.

[44] Jodie L Simpson, Joshua Daly, Katherine J Baines, Ian A Yang, John W Upham, Paul N Reynolds, Sandra Hodge, Alan L James, Philip Hugenholtz, Dana Willner, et al. Airway dysbiosis: Haemophilus influenzae and tropheryma in poorly controlled asthma. *European Respiratory Journal*, pages ERJ–00405, 2015.

[45] Daniela J Vollenweider, Harish Jarrett, Claudia A Steurer-Stey, Judith Garcia-Aymerich, and Milo A Puhan. Antibiotics for exacerbations of chronic obstructive pulmonary disease. *Cochrane Database Syst Rev*, 12, 2012.

[46] Richard K Albert, John Connett, William C Bailey, Richard Casaburi, J Allen D Cooper Jr, Gerard J Criner, Jeffrey L Curtis, Mark T Dransfield, MeiLan K Han, Stephen C Lazarus, et al. Azithromycin for prevention of exacerbations of copd. *New England Journal of Medicine*, 365(8):689–698, 2011.

[47] Marc A Sze, James C Hogg, and Don D Sin. Bacterial microbiome of lungs in copd. *Int J Chron Obstruct Pulmon Dis*, 9:229–238, 2014.

[48] KF Chung. Cytokines in chronic obstructive pulmonary disease. *European Respiratory Journal*, 18(34 suppl):50s–59s, 2001.

[49] Peter J Barnes. Immunology of asthma and chronic obstructive pulmonary disease. *Nature Reviews Immunology*, 8(3):183–192, 2008.

[50] Arthur F Gelb, Noe Zamel, and Anita Krishnan. Physiologic similarities and differences between asthma and chronic obstructive pulmonary disease. *Current opinion in pulmonary medicine*, 14(1):24–30, 2008.

[51] Jean Bousquet, Pascal Chanez, Jean Yves Lacoste, Gilbert Barnéon, Nouchine Ghavanian, Ingrid Enander, Per Venge, Staffan Ahlstedt, Joelle Simony-Lafontaine, Philippe Godard, et al. Eosinophilic inflammation in asthma. *New England Journal of Medicine*, 323(15):1033–1039, 1990.

[52] WM Thurlbeck. Pathophysiology of chronic obstructive pulmonary disease. *Clinics in chest medicine*, 11(3):389–403, 1990.

[53] ANTONINO DI STEFANO, Armando Capelli, MIRCO LUSUARDI, Piero Balbo, Cinzia Vecchio, Piero Maestrelli, Cristina E Mapp, Leonardo M Fabbri, Claudio F Donner, and Marina Saetta. Severity of airflow limitation is associated with severity

of airway inflammation in smokers. *American journal of respiratory and critical care medicine*, 158(4):1277–1285, 1998.

[54] S Saladin Kenneth and MP Carol. Anatomy and physiology: The unity of form and function, 1998.

[55] A Barry Kay. The role of eosinophils in the pathogenesis of asthma. *Trends in molecular medicine*, 11(4):148–152, 2005.

[56] Martin Raff, Bruce Alberts, Julian Lewis, Alexander Johnson, and Keith Roberts. Molecular biology of the cell 4th edition. 2002.

[57] Hongwei Yao and Irfan Rahman. Current concepts on the role of inflammation in copd and lung cancer. *Current opinion in pharmacology*, 9(4):375–383, 2009.

[58] Parameswaran Nair, Marcia MM Pizzichini, Melanie Kjarsgaard, Mark D Inman, Ann Efthimiadis, Emilio Pizzichini, Frederick E Hargreave, and Paul M O'Byrne. Mepolizumab for prednisone-dependent asthma with sputum eosinophilia. *New England Journal of Medicine*, 360(10):985–993, 2009.

[59] Jennifer Kathleen Quint and Jadwiga Anna Wedzicha. The neutrophil in chronic obstructive pulmonary disease. *Journal of allergy and clinical immunology*, 119(5): 1065–1071, 2007.

[60] Kiyoshi Takatsu and Hiroshi Nakajima. Il-5 and eosinophilia. *Current opinion in immunology*, 20(3):288–294, 2008.

[61] Kasper Hoebe, Edith Janssen, and Bruce Beutler. The interface between innate and adaptive immunity. *Nature immunology*, 5(10):971–974, 2004.

[62] Christopher Brightling, Mike Berry, and Yassine Amrani. Targeting tnf-α: a novel therapeutic approach for asthma. *Journal of Allergy and Clinical Immunology*, 121 (1):5–10, 2008.

[63] S Rietveld, I Van Beest, and WTAM Everaerd. Stress-induced breathlessness in asthma. *Psychological medicine*, 29(06):1359–1366, 1999.

[64] RH Green, CE Brightling, G Woltmann, D Parker, AJ Wardlaw, and ID Pavord. 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, 2002.

[65] Shironjit Saha and Christopher E Brightling. Eosinophilic airway inflammation in copd. *International journal of chronic obstructive pulmonary disease*, 1(1):39, 2006.

[66] Jørgen Vestbo and Stephen Rennard. Chronic obstructive pulmonary disease biomarker (s) for disease activity needed—urgently. *American journal of respiratory and critical care medicine*, 182(7):863–864, 2010.

[67] MeiLan K Han, Alvar Agusti, Peter M Calverley, Bartolome R Celli, Gerard Criner, Jeffrey L Curtis, Leonardo M Fabbri, Jonathan G Goldin, Paul W Jones, William MacNee, et al. Chronic obstructive pulmonary disease phenotypes: the future of copd. *American journal of respiratory and critical care medicine*, 182(5):598–604, 2010.

[68] Prescott G Woodruff. Novel outcomes and end points: biomarkers in chronic obstructive pulmonary disease clinical trials. *Proceedings of the American Thoracic Society*, 8(4):350–355, 2011.

[69] Nirav R Bhakta and Prescott G Woodruff. Human asthma phenotypes: from the clinic, to cytokines, and back again. *Immunological reviews*, 242(1):220–232, 2011.

[70] Christopher E Brightling, William Monteiro, Richard Ward, Debbie Parker, Michael DL Morgan, Andrew J Wardlaw, and Ian D Pavord. Sputum eosinophilia and short-term response to prednisolone in chronic obstructive pulmonary disease: a randomised controlled trial. *The Lancet*, 356(9240):1480–1485, 2000.

[71] R Siva, RH Green, CE Brightling, M Shelley, B Hargadon, S McKenna, W Monteiro, M Berry, D Parker, AJ Wardlaw, et al. Eosinophilic airway inflammation and exacerbations of copd: a randomised controlled trial. *European Respiratory Journal*, 29(5):906–913, 2007.

[72] Pranabashis Haldar, Christopher E Brightling, Beverley Hargadon, Sumit Gupta, William Monteiro, Ana Sousa, Richard P Marshall, Peter Bradding, Ruth H Green, Andrew J Wardlaw, et al. Mepolizumab and exacerbations of refractory eosinophilic asthma. *New England Journal of Medicine*, 360(10):973–984, 2009.

[73] Ian D Pavord, Stephanie Korn, Peter Howarth, Eugene R Bleecker, Roland Buhl, Oliver N Keene, Hector Ortega, and Pascal Chanez. Mepolizumab for severe eosinophilic asthma (dream): a multicentre, double-blind, placebo-controlled trial. *The Lancet*, 380(9842):651–659, 2012.

[74] Parameswaran Nair. Mepolizumab in copd with eosinophilic bronchitis: A randomized clinical trial, 2004. URL `https://www.clinicaltrials.gov/ct2/show/record/NCT01463644?term=Anti+IL-5&rank=11`.

[75] Margaret J Leckie, Anneke ten Brinke, Jamey Khan, Zuzana Diamant, Brian J O'Connor, Christine M Walls, Ashwini K Mathur, Hugh C Cowley, K Fan Chung, Ratko Djukanovic, et al. Effects of an interleukin-5 blocking monoclonal antibody on eosinophils, airway hyper-responsiveness, and the late asthmatic response. *The Lancet*, 356(9248):2144–2148, 2000.

[76] Kian Fan Chung, Sally E Wenzel, Jan L Brozek, Andrew Bush, Mario Castro, Peter J Sterk, Ian M Adcock, Eric D Bateman, Elisabeth H Bel, Eugene R Bleecker, et al. International ers/ats guidelines on definition, evaluation and treatment of severe asthma. *European Respiratory Journal*, pages erj02020–2013, 2013.

[77] Jodie L Simpson, Rodney Scott, Michael J Boyle, and Peter G Gibson. Inflammatory subtypes in asthma: assessment and identification using induced sputum. *Respirology*, 11(1):54–61, 2006.

[78] Pranab Haldar and Ian D Pavord. Noneosinophilic asthma: a distinct clinical and pathologic phenotype. *Journal of Allergy and Clinical Immunology*, 119(5):1043–1052, 2007.

[79] Sally E Wenzel, Lawrence B Schwartz, Esther L Langmack, Janet L Halliday, John B Trudeau, Robyn L Gibbs, and Hong Wei Chu. Evidence that severe asthma can be divided pathologically into two inflammatory subtypes with distinct physiologic and clinical characteristics. *American journal of respiratory and critical care medicine*, 160(3):1001–1008, 1999.

[80] John V Fahya, Kwang Woo Kimb, Jane Liub, and Homer A Bousheya. Prominent neutrophilic inflammation in sputum from subjects with asthma exacerbation. *Journal of Allergy and Clinical Immunology*, 95(4):843–852, 1995.

[81] Prescott G Woodruff, Ramin Khashayar, Stephen C Lazarus, Susan Janson, Pedro Avila, Homer A Boushey, Mark Segal, and John V Fahy. Relationship between airway inflammation, hyperresponsiveness, and obstruction in asthma. *Journal of allergy and clinical immunology*, 108(5):753–758, 2001.

[82] Prescott G Woodruff and John V Fahy. A role for neutrophils in asthma? *The American journal of medicine*, 112(6):498–500, 2002.

[83] Annette T Hastie, Wendy C Moore, Deborah A Meyers, Penny L Vestal, Huashi Li, Stephen P Peters, Eugene R Bleecker, National Heart, et al. Analyses of asthma severity phenotypes and inflammatory proteins in subjects stratified by sputum granulocytes. *Journal of Allergy and Clinical Immunology*, 125(5):1028–1036, 2010.

[84] Ruth H Green, Christopher E Brightling, Susan McKenna, Beverley Hargadon, Debbie Parker, Peter Bradding, Andrew J Wardlaw, and Ian D Pavord. Asthma exacerbations and sputum eosinophil counts: a randomised controlled trial. *The Lancet*, 360(9347):1715–1721, 2002.

[85] J Chlumskỳ, I Striz, M Terl, and J Vondracek. Strategy aimed at reduction of sputum eosinophils decreases exacerbation rate in patients with asthma. *Journal of international medical research*, 34(2):129–139, 2006.

[86] L Jayaram, MM Pizzichini, RJ Cook, LP Boulet, C Lemiere, E Pizzichini, A Cartier, P Hussack, CH Goldsmith, M Laviolette, et al. Determining asthma treatment by monitoring sputum cell counts: effect on exacerbations. *European Respiratory Journal*, 27(3):483–494, 2006.

[87] Pranab Haldar, Ian D Pavord, Dominic E Shaw, Michael A Berry, Michael Thomas, Christopher E Brightling, Andrew J Wardlaw, and Ruth H Green. Cluster analysis and clinical asthma phenotypes. *American journal of respiratory and critical care medicine*, 178(3):218–224, 2008.

[88] Valérie Siroux, Xavier Basagaña, Anne Boudier, Isabelle Pin, Judith Garcia-Aymerich, Aurélien Vesin, Rémy Slama, Déborah Jarvis, Josep M Anto, Francine Kauffmann, et al. Identifying adult asthma phenotypes using a clustering approach. *European Respiratory Journal*, 38(2):310–317, 2011.

[89] Jocelyne Just, Rahele Gouvis-Echraghi, Sarah Rouve, Stephanie Wanin, David Moreau, and Isabella Annesi-Maesano. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. *European Respiratory Journal*, 40(1): 55–60, 2012.

[90] Mona Bafadhel, Susan McKenna, Sarah Terry, Vijay Mistry, Carlene Reid, Pranabashis Haldar, Margaret McCormick, Koirobi Haldar, Tatiana Kebadze, Annelyse Duvoix, et al. Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers. *American journal of respiratory and critical care medicine*, 184(6):662–671, 2011.

[91] Pierre Regis Burgel, JL Paillasseur, D Caillaud, I Tillie-Leblond, P Chanez, R Escamilla, T Perez, P Carré, N Roche, et al. Clinical copd phenotypes: a novel approach using principal component and cluster analyses. *European Respiratory Journal*, 36 (3):531–539, 2010.

[92] Mona Bafadhel, Margaret McCormick, Shiron Saha, S McKenna, M Shelley, B Hargadon, V Mistry, C Reid, D Parker, P Dodson, et al. Profiling of sputum inflammatory mediators in asthma and chronic obstructive pulmonary disease. *Respiration*, 83(1):36, 2012.

[93] Dhananjay Desai and Christopher Brightling. Cytokines and cytokine-specific therapy in asthma. *Advances in clinical chemistry*, 57:59, 2012.

[94] Suresh K Bhavnani, Sundar Victor, William J Calhoun, William W Busse, Eugene Bleecker, Mario Castro, Hyunsu Ju, Regina Pillai, Numan Oezguen, Gowtham Bellala, et al. How cytokines co-occur across asthma patients: from bipartite network analysis to a molecular-based classification. *Journal of biomedical informatics*, 44: S24–S30, 2011.

[95] A Bhowmik, TAR Seemungal, RJ Sapsford, JL Devalia, and JA Wedzicha. Comparison of spontaneous and induced sputum for investigation of airway inflammation in chronic obstructive pulmonary disease. *Thorax*, 53(11):953–956, 1998.

[96] Ph H Quanjer, GJ Tammeling, JE Cotes, OF Pedersen, R Peslin, and JC Yernault. Lung volumes and forced ventilatory flows. *European Respiratory Journal*, 6(Suppl 16):5–40, 1993.

[97] A Pye, RA Stockley, and SL Hill. Simple method for quantifying viable bacterial numbers in sputum. *Journal of clinical pathology*, 48(8):719–724, 1995.

[98] Mona Bafadhel, Susan McKenna, Sarah Terry, Vijay Mistry, Mitesh Pancholi, Per Venge, David A Lomas, Michael R Barer, Sebastian L Johnston, Ian D Pavord, et al. Blood eosinophils to direct corticosteroid treatment of exacerbations of chronic obstructive pulmonary disease: a randomized placebo-controlled trial. *American journal of respiratory and critical care medicine*, 186(1):48–55, 2012.

[99] CE Brightling, W Monterio, RH Green, D Parker, MDL Morgan, AJ Wardlaw, and ID Pavord. Induced sputum and other outcome measures in chronic obstructive pulmonary disease: safety and repeatability. *Respiratory medicine*, 95(12):999–1002, 2001.

[100] Ian D Pavord, Chris E Brightling, Gerrit Woltmann, and Andrew J Wardlaw. Noneosinophilic cor ticosteroid unresponsive asthma. *The Lancet*, 353(9171):2213–2214, 1999.

[101] Marie Chavent, Vanessa Kuentz, Benoˆıt Liquet, and L Saracco. Clustofvar: An r package for the clustering of variables. *arXiv preprint arXiv:1112.0295*, 2011.

[102] Barbara G Tabachnick, Linda S Fidell, et al. Using multivariate statistics. 2001.

[103] Henry F Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 1960.

[104] Peter Tryfos. Chapter 14: Factor analysis. *Methods for Business Analysis and Forecasting: Text & Cases,. sl: Wiley*, 1998.

[105] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[106] Brian Everitt and Torsten Hothorn. *An introduction to applied multivariate analysis with R*. Springer Science & Business Media, 2011.

[107] Xindong Wu and Vipin Kumar. *The top ten algorithms in data mining*. CRC Press, 2009.

[108] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[109] Ka Yee Yeung, Mario Medvedovic, and Roger E Bumgarner. Clustering gene-expression data with repeated measurements. *Genome biology*, 4(5):R34, 2003.

[110] Allan R Brasier, Sundar Victor, Hyunsu Ju, William W Busse, Douglas Curran-Everett, Eugene Bleecker, Mario Castro, Kian Fan Chung, Benjamin Gaston, Elliot Israel, et al. Predicting intermediate phenotypes in asthma using bronchoalveolar lavage-derived cytokines. *Clinical and translational science*, 3(4):147–157, 2010.

[111] Perry R Hinton, Isabella McMurray, and Charlotte Brownlow. *SPSS explained*. Routledge, 2014.

[112] Julien Claude. *Morphometrics with R*. Springer Science & Business Media, 2008.

[113] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

[114] LP Stata StataCorp. Stata statistical software: Release 13. *College Station, TX: StataCorp LP*, 2013.

[115] Allan R Brasier, Sundar Victor, Gary Boetticher, Hyunsu Ju, Chang Lee, Eugene R Bleecker, Mario Castro, William W Busse, and William J Calhoun. Molecular phenotyping of severe asthma using pattern recognition of bronchoalveolar lavage–derived cytokines. *Journal of Allergy and Clinical Immunology*, 121(1):30–37, 2008.

[116] Jonathan Corren, Robert F Lemanske Jr, Nicola A Hanania, Phillip E Korenblat, Merdad V Parsey, Joseph R Arron, Jeffrey M Harris, Heleen Scheerens, Lawren C Wu, Zheng Su, et al. Lebrikizumab treatment in adults with asthma. *New England Journal of Medicine*, 365(12):1088–1098, 2011.

[117] Edward Piper, Christopher Brightling, Robert Niven, Chad Oh, Raffaella Faggioni, Kwai Poon, Dewei She, Chris Kell, Richard D May, Gregory P Geba, et al. A phase ii placebo-controlled study of tralokinumab in moderate-to-severe asthma. *European Respiratory Journal*, 41(2):330–338, 2013.

[118] Hiroshi Iwamoto, Jing Gao, Jukka Koskela, Vuokko Kinnula, Hideo Kobayashi, Tarja Laitinen, and Witold Mazur. Differences in plasma and sputum biomarkers between copd and copd–asthma overlap. *European Respiratory Journal*, 43(2):421–429, 2014.

[119] Elizabeth J Atkinson and Terry M Therneau. An introduction to recursive partitioning using the rpart routines. *Rochester: Mayo Foundation*, 2011.

[120] Rachna Shah and Carol A Saltoun. Acute severe asthma (status asthmaticus). In *Allergy and Asthma Proceedings*, volume 33, pages S47–S50. OceanSide Publications, Inc, 2012.

[121] Terence AR Seemungal, Gavin C Donaldson, Elizabeth A Paul, Janine C Bestall, Donald J Jeffries, and Jadwiga A Wedzicha. Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 157(5):1418–1422, 1998.

[122] Shawn D Aaron, Katherine L Vandemheen, Paul Hebert, Robert Dales, Ian G Stiell, Jan Ahuja, Garth Dickinson, Robert Brison, Brian H Rowe, Jonathan Dreyer, et al. Outpatient oral prednisone after emergency treatment of chronic obstructive pulmonary disease. *New England Journal of Medicine*, 348(26):2618–2625, 2003.

[123] GC Donaldson, TAR Seemungal, A Bhowmik, and JA Wedzicha. Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease. *Thorax*, 57(10):847–852, 2002.

[124] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.

[125] Mark O Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973.

[126] Robert Harding Whittaker. Vegetation of the siskiyou mountains, oregon and california. *Ecological monographs*, 30(3):279–338, 1960.

[127] Jari Oksanen, F Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R Minchin, RB O'Hara, Gavin L Simpson, P Solymos, MHH Stevens, and H Wagner. Package 'vegan'. *R Packag ver*, 254:20–8, 2013.

[128] Chris Newby, Liam G Heaney, Andrew Menzies-Gow, Rob M Niven, Adel Mansur, Christine Bucknall, Rekha Chaudhuri, John Thompson, Paul Burton, Chris Brightling, et al. Statistical cluster analysis of the british thoracic society severe refractory asthma registry: clinical outcomes and phenotype stability. 2014.

[129] Adrian E Raftery and Nema Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[130] D Michael Titterington, Adrian FM Smith, Udi E Makov, et al. *Statistical analysis of finite mixture distributions*, volume 7. Wiley New York, 1985.

[131] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[132] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[133] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[134] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[135] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[136] Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[137] Alan Miller. *Subset selection in regression*. CRC Press, 2002.

[138] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[139] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[140] Cathy Maugis, Gilles Celeux, and Marie-Laure Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709, 2009.

[141] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 2010.

[142] Manoranjan Dash and Poon Wei Koot. Feature selection for clustering. In *Encyclopedia of database systems*, pages 1119–1125. Springer, 2009.

[143] Luis Talavera. Dependency-based feature selection for clustering symbolic data. *Intelligent Data Analysis*, 4(1):19–28, 2000.

[144] Martin HC Law, Mario AT Figueiredo, and Anil K Jain. Simultaneous feature selection and clustering using mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1154–1166, 2004.

[145] Chris Fraley and Adrian E Raftery. Mclust version 3: an r package for normal mixture modeling and model-based clustering. Technical report, DTIC Document, 2012.

[146] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458): 611–631, 2002.

[147] Pekka Paalanen. Bayesian classification using gaussian mixture model and em estimation: Implementations and comparisons. *Information Technology Project*, 2004.

[148] Małgorzata Charytanowicz, Jerzy Niewczas, Piotr Kulczycki, Piotr A Kowalski, Szymon Łukasik, and Sławomir Żak. Complete gradient clustering algorithm for features analysis of x-ray images. In *Information technologies in biomedicine*, pages 15–24. Springer, 2010.

[149] S Aeberhard, D Coomans, and O De Vel. The classification performance of rda. *Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland, Tech. Rep*, pages 92–01, 1992.

[150] YeongSeog Kim, W Nick Street, and Filippo Menczer. Feature selection in unsupervised learning via evolutionary search. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 365–369. ACM, 2000.

[151] Eduardo R Hruschka and Thiago F Covoes. Feature selection for cluster analysis: an approach based on the simplified silhouette criterion. In *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, volume 1, pages 32–38. IEEE, 2005.

[152] Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, James Demmel, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, S Hammerling, Alan McKenney, et al. *LAPACK Users' guide*, volume 9. Siam, 1999.

[153] Robert G Stirling, ELIZABETH LJ VAN RENSEN, Peter J Barnes, and K Fan Chung. Interleukin-5 induces cd34+ eosinophil progenitor mobilization and eosinophil ccr3 expression in asthma. *American journal of respiratory and critical care medicine*, 164(8):1403–1409, 2001.

[154] David M Essayan, Charity C Fox, Francesca Levi-Schaffer, Rafeul Alam, and Lanny J Rosenwasser. Biologic activities of il-1 and its role in human disease. *Journal of allergy and clinical immunology*, 102(3):344–350, 1998.

[155] Angshu Bhowmik, Terence AR Seemungal, Raymond J Sapsford, and Jadwiga A Wedzicha. Relation of sputum inflammatory markers to symptoms and lung function changes in copd exacerbations. *Thorax*, 55(2):114–120, 2000.

[156] CE Brightling, S Gupta, S Gonem, and S Siddiqui. Lung damage and airway remodelling in severe asthma. *Clinical & Experimental Allergy*, 42(5):638–649, 2012.

[157] Judith Garcia-Aymerich, Federico P Gómez, Marta Benet, Eva Farrero, Xavier Basagaña, Àngel Gayete, Carles Paré, Xavier Freixa, Jaume Ferrer, Antoni Ferrer, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (copd) subtypes. *Thorax*, 66(5):430–437, 2011.

[158] M Weatherall, J Travers, PM Shirtcliffe, SE Marsh, MV Williams, MR Nowitz, S Aldington, and R Beasley. Distinct clinical phenotypes of airways disease defined by cluster analysis. *European Respiratory Journal*, 34(4):812–818, 2009.

[159] Bethan L Barker and Christopher E Brightling. Phenotyping the heterogeneity of chronic obstructive pulmonary disease. *Clinical Science*, 124(6):371–387, 2013.

[160] Lowie EGW Vanfleteren, Janwillem WH Kocks, Ian S Stone, Robab Breyer-Kohansal, Timm Greulich, Donato Lacedonia, Roland Buhl, Leonardo M Fabbri, Ian D Pavord, Neil Barnes, et al. Moving from the oslerian paradigm to the post-genomic era: are asthma and copd outdated terms? *Thorax*, pages thoraxjnl–2013, 2013.

[161] Peter J Barnes. Against the dutch hypothesis: asthma and chronic obstructive pulmonary disease are distinct diseases. *American journal of respiratory and critical care medicine*, 174(3):240–243, 2006.

[162] Y Cao, W Gong, H Zhang, B Liu, B Li, X Wu, X Duan, and J Dong. A comparison of serum and sputum inflammatory mediator profiles in patients with asthma and copd. *Journal of International Medical Research*, 40(6):2231–2242, 2012.

[163] Michael J Brusco and J Dennis Cradit. A variable-selection heuristic for k-means clustering. *Psychometrika*, 66(2):249–270, 2001.

[164] Charles Bouveyron and Camille Brunet-Saumard. Discriminative variable selection for clustering with the sparse fisher-em algorithm. *Computational Statistics*, 29(3-4): 489–513, 2014.

[165] Benhuai Xie, Wei Pan, and Xiaotong Shen. Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, 64(3):921–930, 2008.

[166] Leonard Poon, Nevin L Zhang, Tao Chen, and Yi Wang. Variable selection in model-based clustering: To do or to facilitate. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 887–894, 2010.

[167] Manali Mukherjee, Roma Sehmi, and Parameswaran Nair. Anti-il5 therapy for asthma and beyond. *World Allergy Organization Journal*, 7(1):32, 2014.

[168] Sinae Kim, Mahlet G Tadesse, and Marina Vannucci. Variable selection in clustering via dirichlet process mixture models. *Biometrika*, 93(4):877–893, 2006.

[169] M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[170] Jeffrey L Andrews and Paul D McNicholas. Variable selection for clustering and classification. *Journal of Classification*, 31(2):136–153, 2014.

[171] Pekka Paalanen. Gmmbcovfixer force matrix to be a valid covariance matrix, 2004. URL `http://www2.it.lut.fi/project/gmmbayes/doc/gmmbayestb-v0.3/gmmbayestb-v0.3/`.

[172] Mepolizumab in copd with eosinophilic bronchitis: A randomized clinical trial, 2015. URL `https://www.clinicaltrials.gov/ct2/show/record/NCT01463644?term=Anti+IL-5+in+COPD&rank=1`.

[173] E Monso, J Ruiz, A Rosell, J Manterola, J Fiz, J Morera, and V Ausina. Bacterial infection in chronic obstructive pulmonary disease. a study of stable and exacerbated outpatients using the protected specimen brush. *American journal of respiratory and critical care medicine*, 152(4):1316–1320, 1995.

[174] Matteo Paoletti, Gianna Camiciottoli, Eleonora Meoni, Francesca Bigazzi, Lucia Cestelli, Massimo Pistolesi, and Carlo Marchesi. 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, 2009.

[175] Romain A Pauwels, A Sonia Buist, Peter MA Calverley, Christine R Jenkins, and Suzanne S Hurd. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 163(5), 2014.

[176] Lisa L Harlow. *The essence of multivariate thinking: Basic themes and methods.* Routledge, 2014.

[177] Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1): 115–123, 2008.

[178] Dirkje S Postma, Scott T Weiss, Maarten van den Berge, Huib AM Kerstjens, and Gerard H Koppelman. Revisiting the dutch hypothesis. *Journal of Allergy and Clinical Immunology*, 136(3):521–529, 2015.

[179] KG Tantisira, AA Litonjua, ST Weiss, and AL Fuhlbrigge. Association of body mass with pulmonary function in the childhood asthma management program (camp). *Thorax*, 58(12):1036–1041, 2003.

[180] Brita Stenius-Aarniala, Tuija Poussa, Johanna Kvarnström, Eeva-Liisa Grönlund, Mikko Ylikahri, and Pertti Mustajoki. Immediate and long term effects of weight reduction in obese people with asthma: randomised controlled study. *Bmj*, 320 (7238):827–832, 2000.

[181] RA Stockley, C O'Brien, A Pye, and SL Hill. Relationship of sputum color to nature and outpatient management of acute exacerbations of copd. 2000. *Chest*, 136 (5 Suppl):e30–e30, 2009.

[182] Annemie MWJ Schols, Peter B Soeters, Annemarie MC Dingemans, Rob Mostert, Peter J Frantzen, and Emiel FM Wouters. Prevalence and characteristics of nutritional depletion in patients with stable copd eligible for pulmonary rehabilitation. *American Review of Respiratory Disease*, 147(5):1151–1156, 1993.

[183] Annemie MWJ Schols, JOS Slangen, Lex Volovics, and Emiel FM Wouters. Weight loss is a reversible factor in the prognosis of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 157(6):1791–1797, 1998.

[184] Donald P Tashkin. Is it asthma, copd, or something in between, and does it matter? *Respiratory care*, 57(8):1354–1356, 2012.

[185] Efrossini Dima, Nikoletta Rovina, Christina Gerassimou, Charis Roussos, and Christina Gratziou. Pulmonary function tests, sputum induction, and bronchial provocation tests: diagnostic tools in the challenge of distinguishing asthma and copd phenotypes in clinical practice. *International journal of chronic obstructive pulmonary disease*, 5:287, 2010.

[186] Paula Kauppi, Henna Kupiainen, Ari Lindqvist, Lauri Tammilehto, Maritta Kilpeläinen, Vuokko L Kinnula, Tari Haahtela, and Tarja Laitinen. Overlap syndrome of asthma and copd predicts low quality of life. *Journal of Asthma*, 48(3): 279–285, 2011.

[187] Megan Hardin, Edwin K Silverman, R Graham Barr, Nadia N Hansel, Joyce D Schroeder, Barry J Make, James D Crapo, Craig P Hersh, COPDGene Investigators, et al. The clinical features of the overlap between copd and asthma. *Respir Res*, 12 (1):127, 2011.

[188] Yoshiaki Kitaguchi, Yoshimichi Komatsu, Keisaku Fujimoto, Masayuki Hanaoka, and Keishi Kubo. Sputum eosinophilia can predict responsiveness to inhaled corticosteroid treatment in patients with overlap syndrome of copd and asthma. *International journal of chronic obstructive pulmonary disease*, 7:283, 2012.

[189] Tae-Bum Kim, Yeon Mok Oh, Yoon-Seok Chang, You Sook Cho, An-Soo Jang, Sang-Heon Cho, Byoung Whui Choi, Sang-Do Lee, and Hee-Bom Moon. The reality of an intermediate type between asthma and copd in practice. *Respiratory care*, 57 (8):1248–1253, 2012.

[190] Gitta H Lubke and Bengt Muthén. Investigating population heterogeneity with factor mixture models. *Psychological methods*, 10(1):21, 2005.

[191] Bengt Muthén, Tihomir Asparouhov, and Irene Rebollo. Advances in behavioral genetics modeling using mplus: Applications of factor mixture modeling to twin data. *Twin research and human genetics*, 9(03):313–324, 2006.

[192] Salman Siddiqui, Sherif Gonem, and Andrew J Wardlaw. Advances in the management of severe asthma. In *Seminars in respiratory and critical care medicine*, volume 33, pages 666–684, 2012.

[193] Kelly Wong McGrath, Nikolina Icitovic, Homer A Boushey, Stephen C Lazarus, E Rand Sutherland, Vernon M Chinchilli, and John V Fahy. A large subgroup of mild-to-moderate asthma is persistently noneosinophilic. *American journal of respiratory and critical care medicine*, 185(6):612–619, 2012.

[194] Peter G Gibson, Jodie L Simpson, and Nicholas Saltos. Heterogeneity of airway inflammation in persistent asthma: evidence of neutrophilic inflammation and increased sputum interleukin-8. *CHEST Journal*, 119(5):1329–1336, 2001.

[195] Yoshiko Kaneko, Yohei Yatagai, Hideyasu Yamada, Hiroki Iijima, Hironori Masuko, Tohru Sakamoto, and Nobuyuki Hizawa. The search for common pathways underlying asthma and copd. *International journal of chronic obstructive pulmonary disease*, 8:65, 2013.

[196] Peter J Barnes and Trevor T Hansel. Prospects for new drugs for chronic obstructive pulmonary disease. *The Lancet*, 364(9438):985–996, 2004.

[197] AV Kamath, ID Pavord, PR Ruparelia, and ER Chilvers. Is the neutrophil the key effector cell in severe asthma? *Thorax*, 60(7):529–530, 2005.

[198] Stanley J Szefler, Richard J Martin, Tonya Sharp King, Homer A Boushey, Reuben M Cherniack, Vernon M Chinchilli, Timothy J Craig, Myrna Dolovich, Jeffrey M Drazen, Joanne K Fagan, et al. Significant variability in response to inhaled corticosteroids for persistent asthma. *Journal of Allergy and Clinical Immunology*, 109(3):410–418, 2002.

[199] Richard J Martin, Stanley J Szefler, Tonya S King, Monica Kraft, Homer A Boushey, Vernon M Chinchilli, Timothy J Craig, Emily A DiMango, Aaron Deykin, John V Fahy, et al. The predicting response to inhaled corticosteroid efficacy (price) trial. *Journal of allergy and clinical immunology*, 119(1):73–80, 2007.

[200] Sally E Wenzel. Eosinophils in asthma—closing the loop or opening the door? *New England Journal of Medicine*, 360(10):1026–1028, 2009.

[201] Celeste Porsbjerg, Thomas Kromann Lund, Lars Pedersen, and Vibeke Backer. Inflammatory subtypes in asthma are related to airway hyperresponsiveness to mannitol and exhaled no. *Journal of asthma*, 46(6):606–612, 2009.

[202] Edited V Brusasco, R Crapo, G Viegi, MR Miller, J Hankinson, V Brusasco, F Burgos, R Casaburi, A Coates, P Enright, et al. Standardisation of spirometry. 2005.

[203] Ferdousi Chowdhury, Anthony Williams, and Peter Johnson. Validation and comparison of two multiplex technologies, luminex® and mesoscale discovery, for human cytokine profiling. *Journal of immunological methods*, 340(1):55–64, 2009.

[204] Mona Bafadhel, S Saha, R Siva, M McCormick, W Monteiro, P Rugman, P Dodson, Ian D Pavord, P Newbold, and Christopher E Brightling. Sputum il-5 concentration is associated with a sputum eosinophilia and attenuated by corticosteroid therapy in copd. *Respiration*, 78(3):256–262, 2009.

[205] A Di Stefano, G Turato, Piero Maestrelli, Cristina E Mapp, Maria Paola Ruggieri, Alberto Roggeri, Piera Boschetto, Leonardo M Fabbri, and Marina Saetta. Airflow limitation in chronic bronchitis is associated with t-lymphocyte and macrophage infiltration of the bronchial mucosa. *American journal of respiratory and critical care medicine*, 153(2):629–632, 1996.

[206] Marina Saetta, Antonino Di Stefano, Piero Maestrelli, Alberto Ferraresso, Riccardo Drigo, Alfredo Potena, Adalberto Ciaccia, and Leonardo M Fabbri. Activated t-lymphocytes and macrophages in bronchial mucosa of subjects with chronic bronchitis. *American review of respiratory disease*, 147(2):301–306, 1993.

[207] Terence C O'Shaughnessy, Tareq W Ansari, Neil C Barnes, and Peter K Jeffery. Inflammation in bronchial biopsies of subjects with chronic bronchitis: inverse relationship of cd8+ t lymphocytes with fev1. *American journal of respiratory and critical care medicine*, 155(3):852–857, 1997.

[208] Jean-Yves Lacoste, Jean Bousquet, Pascal Chanez, Thierry Van Vyve, Joelle Simony-Lafontaine, Nadine Lequeu, Patrice Vic, Ingrid Enander, Philippe Godard, et al. Eosinophilic and neutrophilic inflammation in asthma, chronic bronchitis, and chronic obstructive pulmonary disease. *Journal of allergy and clinical immunology*, 92(4):537–548, 1993.

[209] BE Lams, AR Sousa, PJ Rees, and TH Lee. Subepithelial immunopathology of the large airways in smokers with and without chronic obstructive pulmonary disease. *European Respiratory Journal*, 15(3):512–516, 2000.

[210] Frederick E Hargreave and Richard Leigh. Induced sputum, eosinophilic bronchitis, and chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 160(supplement_1):S53–S57, 1999.

[211] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

[212] Ajith Abraham, Aboul-Ella Hassanien, and Vaclav Snášel. *Computational social network analysis: Trends, tools and research advances*. Springer Science & Business Media, 2009.

[213] Markus Hilty, Conor Burke, Helder Pedro, Paul Cardenas, Andy Bush, Cara Bossley, Jane Davies, Aaron Ervine, Len Poulter, Lior Pachter, et al. Disordered microbial communities in asthmatic airways. *PloS one*, 5(1):e8578, 2010.

[214] Weiliang Qiu and Harry Joe. *clusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*, 2015. URL `http://CRAN.R-project.org/package=clusterGeneration`. R package version 1.3.4.