BRN Reviews REVIEW

Usefulness of Biological Clustering Patterns in Chronic Obstructive Pulmonary Disease

Andreas Halner, BA and Mona Bafadhel, PhD, FRCP

Respiratory Medicine Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK

ABSTRACT

Chronic obstructive pulmonary disease (COPD) affects millions of people worldwide. It is now clear that COPD is heterogeneous, different components of the disease being present in different patients. Yet, the diversity of COPD pathophysiology, severity and how this relates to disease prognosis and treatment outcomes is far from understood. In order to address this, mathematical techniques such as cluster analysis have been employed to identify subgroups or clusters of COPD patients with differing disease attribute profiles. However, significant methodological shortcomings call into question the validity of the COPD clusters identified in such studies. Furthermore, few published studies relate COPD clusters to underlying disease mechanisms and treatment outcomes. Where this has been addressed, progress has particularly been made for patients with an eosinophilic-predominant profile. In order to maximise the usefulness of COPD cluster analysis studies, we propose that future studies must implement more stringent methodologies and focus on COPD inflammatory biology. **(BRN Rev. 2019;5(3):201-214)**

Corresponding author: Mona Bafadhel, mona.bafadhel@ndm.ox.ac.uk

Key words: Chronic obstructive pulmonary disease. Cluster analysis. Machine learning.

Correspondence to:
Mona Bafadhel, PhD, FRCP
Respiratory Medicine Unit, Nuffield Department of Clinical Medicine,
NDM Research Building, Old Road Campus University of Oxford,
Oxford OX3 7FZ, United Kingdom
E-mail: mona.bafadhel@ndm.ox.ac.uk

Received in original form: 07-03-2019
Accepted in final form: 03-04-2019
DOI: 10.23866/BRNRev:2018-0016

INTRODUCTION

The worldwide prevalence of chronic obstructive pulmonary disease (COPD) has been estimated as 251 million people and is now the third leading cause of death worldwide[1-3]. The disease is characterised by inflammation of the airways leading to a variety of disease features including chronic bronchitis, in which there is thickening of the bronchiolar walls and regular sputum production, and emphysema, in which alveoli are destroyed[4]. Both of these disease characteristics lead to increased airways resistance and loss of elastic recoil of the airways, resulting in expiratory flow limitation which ultimately leads to the dyspnoea suffered by patients[5,6]. Chronic obstructive pulmonary disease is a progressive condition, punctuated by periods of acute respiratory symptom worsening known as exacerbations[4]. Tobacco smoking is the main risk factor for COPD, directly damaging airway epithelium with associated recruitment of inflammatory cells[5,7].

COPD has long been recognised as a "heterogenous" disease, meaning not all of its components are present in all patients or at all time points in a given patient[8]. Recognition of COPD heterogeneity was first made in 1955 when Dornhorst et al.[9] identified two types of patients - emphysematous patients with dyspnoea and muscle wasting ("pink puffers") and chronic bronchitic patients with cyanosis and right heart failure ("blue bloaters"). For a long time, a forced expiratory volume in one second (FEV_1)-centric view of COPD prevailed and initial classifications of COPD severity used four categories, from 1 to 4 in order of increased severity[10,11]. While spirometric measurements confirming post-bronchodilator FEV_1/forced vital capacity (FVC) < 0.7 remain key for COPD diagnosis today, it is now appreciated that patients' exacerbation frequency history and dyspnoea severity are better predictors than FEV_1 of future exacerbation risk and mortality, respectively[12,13]. However, a number of facets of diversity in COPD severity are not adequately accounted for[4,14]. For example, it is acknowledged that COPD pathophysiology involves a diverse range and degree of inflammatory profiles, yet how COPD inflammatory profiles relate to disease progression and treatment outcomes remains poorly understood[5]. Similarly, it is known that comorbidities are common in COPD patients and adversely impact on mortality but an integrated understanding of how comorbidities relate to underlying pathophysiological and clinical features of COPD in different patients is lacking[4,15,16].

The lack of insight into COPD heterogeneity has hindered the development of treatments which significantly alter the course of disease; current pharmacological treatment for COPD, typically consisting of bronchodilator therapy and inhaled corticosteroids (ICS), are often empiric in nature, with limited consideration of the diversity of COPD severity or associated underlying biological mechanisms[17]. Furthermore, current treatments improve symptoms but do not markedly alter the course of the disease, have clinical deteriorations during treatment and an associated side effect profile[18]. This emphasises the need to identify subgroups of COPD patients for whom the benefits of existing treatment significantly outweigh the risks, as well as to devise new treatments targeting disease mechanisms specific to particular subgroups of patients[4,5]. Research has hence sought

to identify different COPD "phenotypes" which may represent unique prognostic and therapeutic subgroups in the COPD population. Currently, a COPD clinical phenotype is defined as *"a single or combination of disease attributes that describe differences between individuals with COPD as they relate to clinically meaningful outcomes (symptoms, exacerbations, response to therapy, rate of disease progression, or death)"*[19]. This has been accompanied by the search for "endotypes", namely, *"subtype(s) of disease defined functionally and pathologically by a molecular mechanism or by treatment response"*[20]. The large number of variables to be taken into account for meaningful COPD phenotyping and endotyping has prompted the use of an unsupervised machine learning technique known as cluster analysis (CA) in airways disease heterogeneity research[21]. The aim of this review is to critically evaluate the usefulness of attempts to identify COPD clusters and how these relate to disease prognosis and pathophysiology to inform optimisation of treatment outcome.

CLUSTER ANALYSIS – A METHOD FOR SUBGROUPING CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Cluster analysis seeks to organise data points representing, for example, individual patients, from a heterogeneous population into subgroups or "clusters" of relative homogeneity[22]; it also enables the identification of clusters for high dimensional data which cannot be visualised, hence the usefulness of CA in the context of recognising the diversity of COPD subgroups and for the purpose of phenotyping and endotyping. Where there are a very large number of variables to consider, as is the case in COPD heterogeneity research, a technique known as principal component analysis (PCA) for continuous variables and multiple correspondence analysis (MCA) for categorical variables, is commonly performed before CA to identify a reduced number of new, independent variables or dimensions for use as input for CA[16,23]. Figure 1 illustrates the concept of clustering.

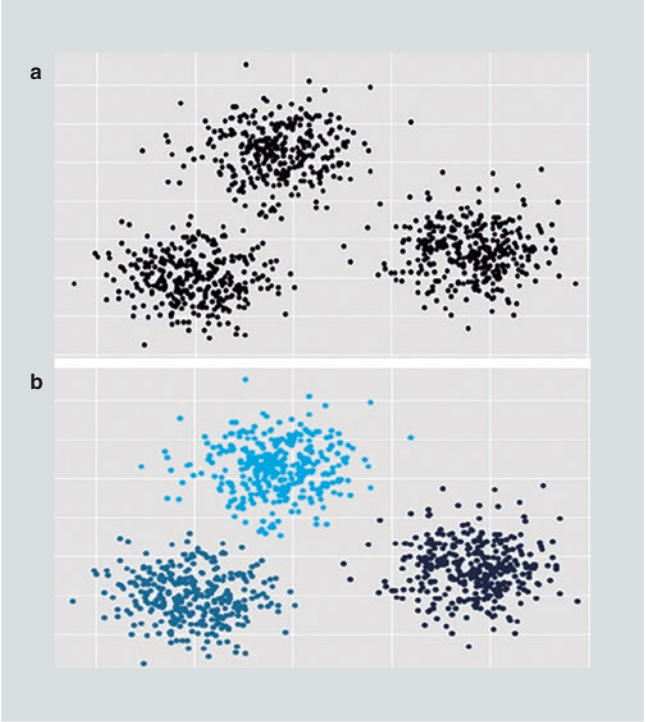

FIGURE 1. Illustration of cluster analysis. **a**. A heterogeneous population, featuring three apparent subgroups, is shown in two dimensions. **b**. K-means algorithm enables identification of the subgroups (highlighted light blue, dark blue and black).

There are a number of ways in which a cluster can be mathematically defined; for any given dataset, different cluster-finding algorithms may identify slightly different clusters. Each cluster-finding algorithm is associated with advantages and limitations. Table 1 summarises some of these considerations for two of the most commonly used cluster-finding algorithms in COPD heterogeneity research.

Table 1. A summary of the procedure of and considerations to be made for K-means and Ward's methods of clustering

Clustering Technique	K-Means (Partitional)[22,24,25]	Ward's Method (Agglomerative Hierarchical)[26]
How it works	i) *K* initial centroids are chosen, where *K* is the pre-specified number of clusters. ii) Each data point is assigned to the nearest centroid, as determined by a proximity measure (e.g., Euclidean distance). iii) The position of each centroid is updated based on the data points belonging to each centroid's cluster, and the data points are then reassigned to the new nearest centroid. iv) These steps are repeated until the centroids remain the same. The final grouping of data points around the centroids constitutes the final clusters.	i) Each data point initially represents a cluster. ii) Clusters which are most similar to each other as determined by a proximity measure (e.g., Euclidean distance) are merged with each other. iii) This step is repeated until only one cluster remains. iv) A dendrogram enables geometrical interpretation of the tree of clusters.
Strengths	Computationally efficient.	Number of clusters does not need to be pre-specified by the user since all cluster-subcluster relationships are evident from dendrogram representation.
Limitations	i) Number of clusters must be pre-specified by the user. ii) Highly sensitive to position of initial centroids in step 1 of algorithm. iii) Highly sensitive to outliers. iv) Performs poorly if the true clusters are non-spherical, or if clusters vary considerably in size or density.	i) Although the number of clusters need not be pre-specified by the user in order to run the algorithm, a method is required for deciding on the most meaningful clusters to extract from the dendrogram. ii) Computationally expensive. iii) Cluster merging decisions in the early steps of the algorithm to maximise local optimization may lead to reduced optimization in downstream merging decisions, hence reducing global optimization.
Possible solutions to limitations	A number of cluster evaluation methods exist for deciding how many clusters from 2 to n, where n is the number of data points, are suitable e.g., pseudo F statistic measures the ratio of between cluster variance to within-cluster variance (Calinski and Harabasz stopping rule). Bisecting K-means algorithm is less susceptible to initial centroid positions. Outliers can be removed before clustering or identified in a post-cluster processing step.	A number of cluster evaluation methods exist for deciding how many clusters from 2 to n, where n is the number of data points, are suitable e.g., pseudo F statistic measures the ratio of between cluster variance to within-cluster variance (Calinski and Harabasz stopping rule).

There is no single clustering algorithm which performs best in all scenarios; this follows from the "no free lunch" theorem for unsupervised optimization problems, including CA, which states that "any two algorithms are equivalent when their performance is averaged across all possible problems"[27,28]. Thus, when performing COPD CA, the algorithm should be selected based on the nature of the data to be clustered and the ease of use of the algorithm for the particular context. Necessary steps should then be taken to address the limitations associated with the chosen algorithm. The usefulness of attempts to identify COPD clusters will now be evaluated.

COPD CLUSTERS – THE FINDINGS

Over the last decade there have been a multitude of studies using CA to identify COPD subgroups. The holy grail of COPD heterogeneity research is the identification of COPD subgroups which provide prognostic power superior to that of the current Global Initiative for Chronic Obstructive Lung Disease

Table 2. Summary of chronic obstructive pulmonary disease (COPD) phenotyping and endotyping study findings

Study	Summary of Findings
Burgel et al. (2010)[29]	Multicentre. Clinical variables at stable state. Finding of four clusters which differed in terms of patient age, severity of airflow limitation and comorbidity.
Garcia-Aymerich et al. (2011)[30]	Multicentre. Clinical variables at exacerbation. Finding of three clusters which differed in terms of severity of airflow limitation and comorbidity. Longitudinal follow-up revealed cluster differences in hospitalisation rate and all-cause mortality.
Bafadhel et al. (2011)[31]	Single-centre. Biological variables at exacerbation. Finding of four clusters which differed in inflammatory profile.
Burgel et al. (2012)[32]	Multicentre. Clinical variables at stable state. Finding of three clusters which differed in terms of patient age, severity of airflow limitation and comorbidity. Longitudinal follow-up revealed cluster differences in all-cause mortality.
Castaldi et al. (2014)[33]	Multicentre. Clinical variables at stable state. Finding of four clusters which differed in degree of airflow obstruction and emphysema as well as COPD-associated gene variants.
Rennard et al. (2015)[34]	Multicentre. Clinical and biological variables at stable state. Finding of five clusters which differed in degree of airways disease, inflammation and comorbidity. Longitudinal follow-up revealed cluster differences in all-cause mortality and time to first exacerbation.
Esteban et al. (2016)[35]	Multicentre. Clinical variables at stable state. Finding of four clusters which differed in terms of dyspnoea, lung function, health-related quality of life and comorbidity. Longitudinal follow-up revealed cluster differences in mortality. Stability of clusters over time was confirmed.
Chang et al. (2016)[36]	Multicentre. Biological variables at stable state. Finding of four clusters which differed in terms of lung function impairment and inflammatory profiles.
Burgel et al. (2017)[37]	Multicentre. Clinical variables at stable state and exacerbation. Finding of five clusters which differed in terms of patient age, severity of airflow limitation and comorbidity. Longitudinal follow-up revealed cluster differences in all-cause mortality.
Zarei et al. (2017)[38]	Multicentre. Biological variables at stable state. Finding of three clusters which differed in terms of degree of airways disease and disease-related quality of life.

(GOLD) classification system and/or correspond to distinct pathophysiological subgroups within the COPD population, enabling the development of treatments specific to the disease-causing mechanism in individual patients[19]. We have selected for critical evaluation examples only of studies which either a) compare identified COPD subgroups against currently used measures such as exacerbation frequency history and dyspnoea severity for predicting long term clinical outcomes, b) compare identified COPD subgroups in terms of relevant clinical outcome measures longitudinally and c) have focused on sputum- and/or serum-based biological COPD subgroups which may correspond to potential aetiology and inflammation. A summary of the relevant studies is provided in table 2 (for more detail, refer to Appendix 1) and figure 2.

Altogether, the studies by Burgel et al. (2010; 2012; 2017)[29,32,37] and Garcia-Aymerich et al. (2011)[30] identified clusters corresponding to patients with severe respiratory symptoms and high prevalence of comorbidities, severe respiratory symptoms and low prevalence of comorbidities, milder respiratory symptoms and high prevalence of comorbidities, and milder respiratory symptoms with few comorbidities. In each of these studies, apart from Garcia-Aymerich et al.[30], the clusters with high prevalence of comorbidities were also the clusters with higher patient age[29,32,37]. It is known that the prevalence of comorbidities

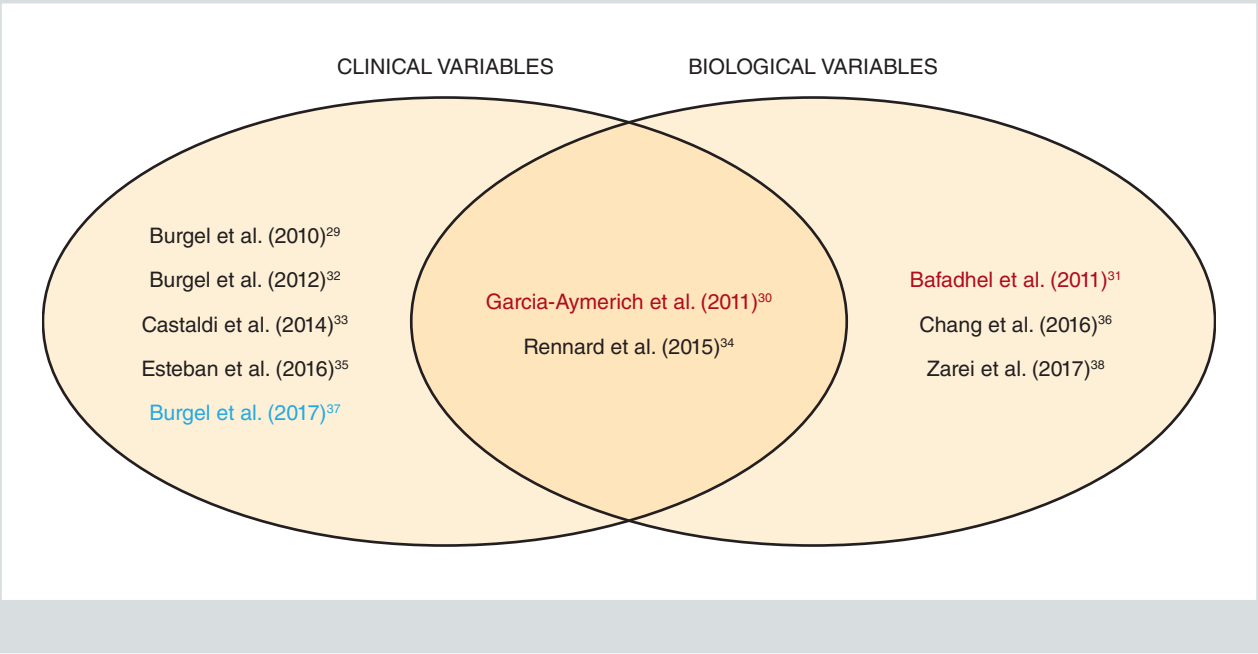

Figure 2. Summary of chronic obstructive pulmonary disease cluster analysis studies in terms of whether clinical and/or biological variables were used (studies performing cluster analysis for patients at stable state are shown in black; studies performing cluster analysis for patients at exacerbation are shown in red; studies performing cluster analysis for a combination of patients at stable state and at exacerbation are shown in blue).

increases with age in the general population[4], which may explain the aforementioned trend. In all four studies, longitudinal follow-up of patients was performed to compare mortality rates between the clusters; in all cases, the cluster with the most severe respiratory symptoms – in terms of dyspnoea and airflow obstruction – was reported as having the highest mortality rate[29,30,27,39]. However, a number of methodological limitations are noted. Garcia-Aymerich et al.[30] and Burgel et al.[29,32,37] included FEV_1- and dyspnoea severity-related measures among the variables used for CA (either directly or in order to generate components or factors which were then used in CA) but did not subsequently adjust for FEV_1 and dyspnoea differences between clusters when comparing clusters for all-cause mortality. The same limitation applies to the studies performed by Rennard et al.[34] and Esteban et al.[36]. This can render it improbable to conclude that the clusters identified provide greater prognostic ability than current standards.

While there are no strict sample size rules for cluster analysis, research in the field of cluster and latent class analysis suggests a much greater sample size than number of variables should be used; otherwise the number of theoretically possible cluster output solutions may be unacceptably high which renders cluster solutions unreliable[40,41]. One study in table 2 violated this assumption[30]. Other kinds of methodological errors feature in some studies described in table 2. For example, in one case PCA and MCA were separately applied on continuous and categorical

variables, respectively, in order to reduce the variables to obtain suitable components for CA[31]. In so doing, there can be a failure to account for correlation between continuous and categorical variables. This may bias the importance of selected components in accounting for COPD severity variation which would render the clusters inaccurate. A technique known as factor analysis of mixed variables (FAMD) can be used to simultaneously reduce the continuous and categorical variables so that the factors obtained accurately explain the variation within the data based on the entire set of variables rather than just a proportion[42]. The variables with the highest loading for the FAMD-derived unbiased factors can then be used as input for CA, allowing the clusters obtained to more accurately reflect the true pattern of COPD patient data.

The Castaldi et al.[33] and Bafadhel et al.[31] studies provide additional methods attempting to identify COPD phenotypes and endotypes respectively. The patient cohort in the Castaldi et al.[33] study was the largest of all studies in table 2. Unlike other studies, Castaldi et al. assessed various CA models for the quality of clusters in order to choose the most appropriate CA model. Statistically significant differences between clusters in terms of relevant clinical outcome measures including dyspnoea and exacerbation frequency were then demonstrated after differences in GOLD stage membership between clusters had been adjusted for[33]. However, similar to all the cluster analyses discussed thus far, the study by Castaldi et al. was cross-sectional and inferred little information about underlying mechanisms. Only Bafadhel et al.[31] have addressed underlying biology in cluster analysis, focusing exclusively on biomarkers. Importantly, the biological clusters based on sputum mediators were related to serum mediators as well as differential blood and sputum neutrophil and eosinophil counts and sputum microbiology but without any clinical differences such as in FEV_1 or exacerbation rate being demonstrated between each biological cluster[31]. This is an important realisation since it may confirm that clinical symptoms alone are not sufficient to fully characterise COPD[31]. Furthermore, in patients with multiple exacerbations during the one-year study, biological clusters were repeatable[31]. Few other studies can be found in which COPD CA has been performed solely on sputum- and serum-based data. In cases where such work has been conducted, for example, by Chang et al.[36] and Zarei et al.[38], only peripheral blood-based data were used for CA without validating whether they are related to pulmonary inflammation.

Despite the strengths of the Bafadhel et al.[31] study compared with other COPD CA studies, there is a paucity of data showing if patients remain in the same cluster over the long term as the disease progresses. These are important considerations - if the inflammatory profile underlying COPD changes in a given patient with time, the kind of treatment needed to target the underlying mechanisms may also need to be adapted with time. If longitudinally stable clusters based on biological data such as sputum and serum inflammatory markers are found to exist in the COPD population, it would be useful to train an algorithm to assign patients to appropriate clusters to identify to which COPD pathophysiological, prognostic and/or therapeutic subgroups individual patients belong.

TO WHAT EXTENT DO COPD-CLUSTERS REPRESENT DISTINCT, REPRODUCIBLE SUBGROUPS WHICH ADEQUATELY REFLECT COPD HETEROGENEITY?

Critical evaluation reveals that no COPD CA studies directly determined whether the patient data used for CA exhibits natural groupings of patients or whether the clusters found in these studies are merely an artefact of the cluster-finding methodology. The majority of cluster algorithms, including k-means and Ward's method, find clusters even if the data display no natural clusters; this is illustrated in figure 3.

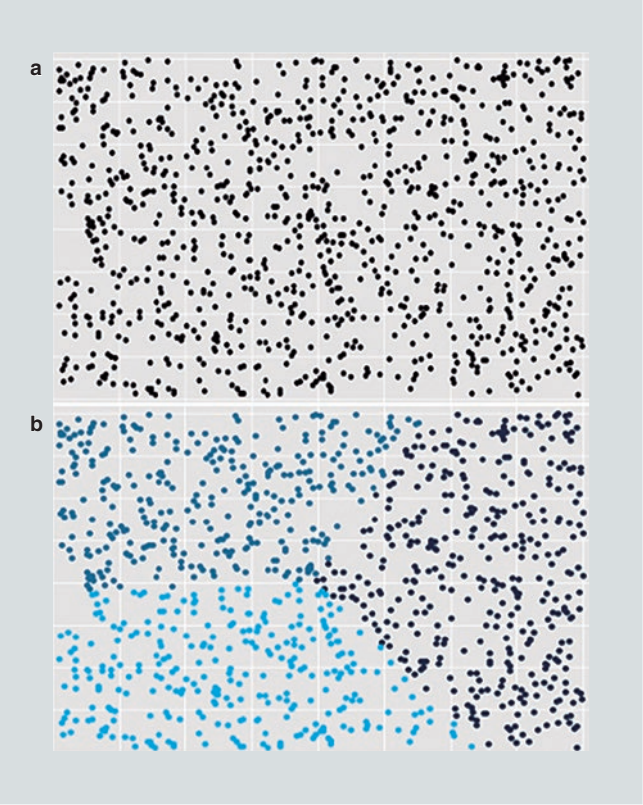

Figure 3. Illustration of cluster identification in absence of natural structure in data. **a**. Randomly distributed data, shown in two dimensions. **b**. Applying the K-means clustering algorithm leads to cluster identification (highlighted light blue, dark blue and black) despite the data exhibiting no natural structure.

Techniques such as pseudo-f statistics and normalised mutual information (NMI) used in COPD CA studies to determine cluster number do not suffice for assessing cluster structure since they assume that at least two clusters are present[22,23]. This emphasises the importance of visualising clusters, where possible, as was done in the Bafadhel et al.[31] study (see Fig. 4).

For studies in which cluster visualisation is unfeasible, for example, where high dimensional data has not been reduced prior to CA, it is important that mathematical approaches are used to determine whether the data possesses cluster structure. Cluster structure can be assessed using spatial randomness tests such as the Hopkins statistic[43,44]. Since the purpose of COPD CA studies is to identify subgroups of patients with unique prognostic, pathophysiological and therapeutic characteristics, absence of natural cluster structure in the data defeats the goal of performing CA even if apparent clusters are found by the algorithm. In keeping with this concern, a recent study assessing 17,146 patients from 10 independent cohorts found limited reproducibility of the COPD clusters across different cohorts[45]. The PCA plots of the pooled patient data from different COPD clustering studies revealed the data was organised as a continuum rather than discrete clusters[45]. However, the cluster reproducibility study did not include biomarker data.

Another important consideration when performing CA is the extent to which the patients included in the analysis reflect the COPD population at large; CA will only be

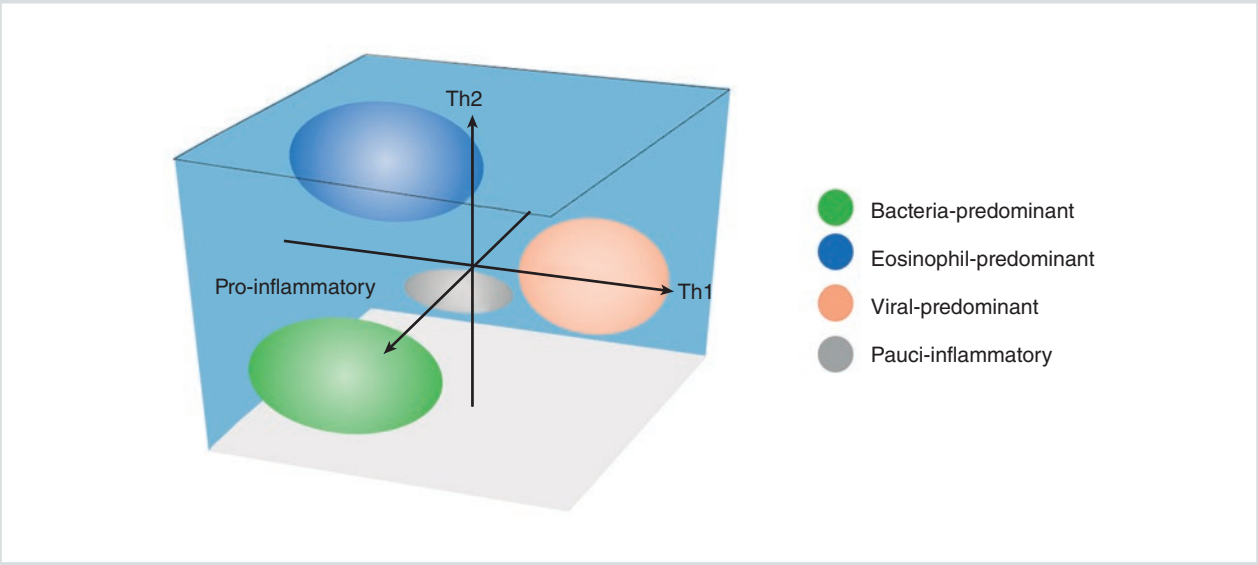

Figure 4. Biological COPD exacerbation clusters shown as ellipsoids in 3-dimensional space using 3 factors (proinflammatory, Th1 and Th2) as coordinate axes. There are 4 clear clusters (bacteria-predominant, eosinophil-predominant, viral-predominant and pauci-inflammatory) *(adapted with permission from Bafadhel M et al.[31]).*
COPD: chronic obstructive pulmonary disease; Th: T helper.

able to identify subgroups of COPD patients present within the patient sample included in the analysis. Hence, the insights which can be drawn from the CA studies of table 2 depend on the extent to which these studies have used representative samples of the COPD population. The majority of COPD CA studies include patients across all four GOLD severity categories. However, previous studies in which spirometry and symptom questionnaires were performed for a random sample of the general population have shown that only a minority of individuals with COPD as determined by spirometry and symptom scores have a previous COPD diagnosis[46,47]; in other words, there is a high rate of under-diagnosis for COPD, particularly for patients with less severe symptoms[46,47]. Furthermore, in recent years it has become increasingly apparent that some patients have disease attributes best described as characterising an overlap of asthma and COPD[48]. Such patients are often neglected from COPD studies that have investigated phenotypes and endotypes using CA. Table 3 (for more detail, see Appendix 2) shows two examples of studies (from the same authors) attempting to take this into account by identifying inflammatory profile-based clusters for a mixed sample of asthma and COPD patients at stable state and during exacerbations.

The field of COPD-CA would benefit from protocols in which a random sample of the general population is first selected after which spirometric measurements and symptom questionnaires are performed to identify the broad spectrum of COPD patients, some of whom overlap with asthma, in the

Table 3. Examples of cluster analysis studies including both asthma and chronic obstructive pulmonary disease (COPD) patients

Study	Summary of Findings
Ghebre et al. (2015)[49]	Single-centre. Biological variables at stable state. Finding of three clusters which differed in inflammatory profile, clinical symptoms and proportion of COPD patients versus asthma patients.
Ghebre et al. (2018)[50]	Single-centre. Biological variables at exacerbation. Finding of three clusters which differed in inflammatory profile, clinical symptoms and proportion of COPD patients versus asthma patients.

population. Clusters obtained from such studies would then be more informative regarding the diversity of COPD subgroups in the population.

USING INSIGHTS FROM COPD-CLUSTER ANALYSIS STUDIES TO OPTIMISE TREATMENT OUTCOMES AND INFORM NEW TREATMENTS

As has been shown, the vast majority of COPD CA studies are hampered by methodological shortcomings and have focused on clinical features of disease without considering endotypes. These limitations may render such studies uninformative when trying to relate COPD clusters to outcomes for current treatment strategies and when devising new treatments aimed at subgroups within the COPD population. However, COPD CA has immense potential for optimising treatment outcomes if performed with appropriate methodology and appropriate focus on endotypes. The lack of recognition of the diversity of COPD may underlie the mixed outcomes of clinical trials for new treatments aiming to target COPD pathophysiological mechanisms. For example, a recent phase 2 trial tested the efficacy of an antagonist (MK-7123) of the cytokine receptor CXCR2 in COPD patients[51]. Since neutrophils are believed to be the key inflammatory cell type in COPD pathophysiology, it was believed that reducing neutrophil chemotaxis by antagonising CXCR2 would significantly improve lung function in a wide spectrum of COPD patients included in the study[51-53]. However, statistically significant improvement in lung function and exacerbations was seen only in COPD patients who were current smokers[51]. The authors concluded that this may be the result of smokers having ongoing exposure-induced neutrophil recruitment to their airways and hence exhibiting more neutrophilia than ex-smokers, rendering the neutrophil chemotaxis-reducing MK-7123 treatment more effective in current smokers[51]. However, the authors did not assess differences in sputum neutrophil levels between current and ex-smokers at the time of starting the MK-7123 treatment and it is possible that there is a subgroup of current and ex-smokers with high sputum neutrophil levels who would benefit from MK-7123 treatment. This would be in keeping with the findings by Bafadhel et al.[31] – the bacteria-predominant and pauci-inflammatory clusters exhibited significantly higher sputum neutrophil levels than patients in the remaining clusters, independent of smoking history. Thus, even in cases where a particular pathophysiological mechanism is common in the COPD patient population, a biological CA-based approach to COPD endotyping may be crucial in revealing subgroups likely

to benefit most from a particular treatment approach.

The bacteria-predominant, viral-predominant and eosinophil-predominant biological exacerbation clusters identified by Bafadhel et al.[31], characterising 55%, 29% and 28% of exacerbations respectively in this study, point to more specific pathophysiological mechanisms which may underlie exacerbations in particular subgroups of patients. Antibiotic use is already recommended for COPD patients in the treatment of exacerbations if the patient displays purulent sputum, a strategy which has been shown to reduce short term mortality and treatment failure[4,54]. However, the use of antibiotics in the long-term management of COPD in patients with sputum purulence in stable state to improve long term outcomes remains controversial, largely due to concerns regarding antibiotic resistance[54]. Treatment targeting subgroups of COPD patients with virus-associated exacerbations is also likely to be difficult due to the limited availability of suitable antiviral agents, particularly for rhinovirus which is the most common virus associated with exacerbations[31,55,56]. In contrast, numerous potentially suitable approaches are available for targeting the eosinophil-predominant COPD cluster to optimise COPD treatment outcomes. These will now be briefly discussed as an example of the potential usefulness of COPD CA approaches.

EOSINOPHILIC COPD-CLUSTER – A CASE STUDY IN PERSONALISED COPD TREATMENT

Although the role of eosinophils in the pathogenesis of COPD is not understood, it has been shown that a raised peripheral blood eosinophil count, a sensitive and specific biomarker for eosinophilic airway inflammation, is associated with poorer COPD clinical outcomes, including increased risk of exacerbations and COPD-specific mortality[31,57-59]. There have been attempts to directly target eosinophilic airway inflammation in COPD by the use of monoclonal antibodies against the interleukin (IL)-5 receptor, which plays an important role in signalling pathways facilitating eosinophil activation[60,61]. So far these studies have been unsuccessful in improving lung function and other relevant clinical outcome measures but since these studies were underpowered for detecting beneficial effects of the anti-IL-5, future investigations using larger patient cohorts are essential[60,61]. Another use of the eosinophilic COPD cluster, supported by a plethora of evidence, is its use in informing treatment outcomes - predicting which COPD patients respond best to corticosteroid treatment. Post-hoc analyses of previous clinical trials have shown patients with a raised peripheral blood eosinophil count to respond more effectively to ICS or oral glucocorticoids, in terms of clinical measures such as lung function and dyspnoea[62,63]. A recent meta-analysis of three randomised control trials showed that COPD patients with a blood eosinophil count ≥ 2% and who did not receive prednisolone, an oral corticosteroid, had a significantly greater treatment failure rate (defined as retreatment, hospitalisation or death within 90 days of randomisation) than patients with a blood eosinophil count ≥ 2% who received prednisolone[64]. However, a caveat to such studies is the use of cut-off points in order to assign patients to a low or high blood eosinophil group, as patients may fluctuate around this cut-off value as seen in

analysis of the Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) cohort[65]. This limitation can be overcome by modelling the eosinophil count as a continuous variable to identify predictors of response to ICS[66]. In the first study of its kind, in a post hoc analysis the authors found that there was a non-linear increase in annual exacerbation rate reduction with increasing blood eosinophil count for patients receiving the ICS/long-acting β_2-agonists (LABA) combination compared with patients receiving LABA alone[66]. Prospective studies are warranted to determine the blood eosinophil count and suitable corresponding corticosteroid dose for a minimal clinically important difference for treatment outcome in the general COPD population. Nevertheless, the Bafadhel et al.[66] study strongly suggests that the severity of COPD eosinophilia, as indicated by peripheral blood eosinophil count, has the potential to guide corticosteroid treatment for COPD patients to maximise treatment outcome for the relevant patient subgroups.

CONCLUSIONS AND FUTURE DIRECTIONS

Identifying COPD clusters corresponding to different patterns of disease has been widely regarded as the holy grail of facilitating personalised COPD management strategies as an alternative to the current suboptimal one-size-fits-all approach. However, we believe that methods for cluster analysis need to be reviewed and conducted with accuracy to reduce statistical errors. The majority of studies focus solely on clinical variables but make no attempt to relate the apparent COPD clusters to treatment outcomes or underlying mechanisms, which would have the potential for maximising COPD treatment outcomes, as exemplified by the use of peripheral blood eosinophil count to guide corticosteroid treatment. In keeping with these considerations, we conclude that the attention on characterising the diversity of COPD heterogeneity must now focus on COPD inflammatory profiles, and the reproducibility of these inflammatory profile COPD clusters over time and between different studies must then be assessed. The potentially different pathophysiological mechanisms underlying different inflammatory profile COPD clusters can be explored and treatments developed to directly target these disease mechanisms. This will enable COPD treatment outcomes to be truly maximised.

DISCLOSURES

Mr. Halner has nothing to disclose. Dr. Bafadhel reports grants from AstraZeneca, and personal fees from AstraZeneca, Boehringher Ingelheim, Cheisi and GSK outside the submitted work.

REFERENCES

1. Vos T, Abajobir AA, Abbafati C et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390:1211–59.
2. Abajobir AA, Abbafati C, Abbas KM et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390:1151–1210.
3. The top 10 causes of death [Internet]. [cited 2019 Mar 4]. Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.
4. Vogelmeier CF, Criner GJ, Martinez FJ et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. Eur Respir J. 2017;49.
5. Decramer M, Janssens W, Miravitlles M. Chronic obstructive pulmonary disease. Lancet. 2012;379:1341–51.

6. O'Donnell DE. Hyperinflation, Dyspnea, and Exercise Intolerance in Chronic Obstructive Pulmonary Disease. Proc Am Thorac Soc. 2006;3:180–4.
7. Kohansal R, Martinez-Camblor P, Agustí A, Sonia Buist A, Mannino DM, Soriano JB. The natural history of chronic airflow obstruction revisited: An analysis of the Framingham Offspring Cohort. Am J Respir Crit Care Med. 2009;180:3–10.
8. Agusti A, Gea J, Faner R. Biomarkers, the control panel and personalized COPD medicine. Respirology. 2016;21:24–33.
9. Dornhorst AC. Respiratory Insufficiency. Lancet. 1955;265:1185–7.
10. Pauwels RA, Buist AS, Calverley PMA, Jenkins CR, Hurd SS. NHLBI/WHO Workshop Summary Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop Summary. 2001;163:1256–76.
11. Rabe KF, Hurd S, Anzueto A et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med. 2007;176:532–55.
12. Hurst JR, Vestbo J, Anzueto A et al. Susceptibility to Exacerbation in Chronic Obstructive Pulmonary Disease. N Engl J Med. 2010;363:1128–38.
13. Nishimura K, Izumi T, Tsukino M, Oga T. Dyspnea is a better predictor of 5-year survival than airway obstruction in patients with COPD. Chest. 2002;121:1434–40.
14. Vestbo J, Hurd SS, Agustí AG et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease GOLD executive summary. Am J Respir Crit Care Med. 2013;187:347–65.
15. Fabbri LM, Luppi F, Beghé B, Rabe KF. Complex chronic comorbidities of COPD. Eur Respir J. 2008;31:204–12.
16. Burgel P-R, Paillasseur J-L, Roche N. Identification of Clinical Phenotypes Using Cluster Analyses in COPD Patients with Multiple Comorbidities. Biomed Res Int. 2014;2014.
17. Agusti A. The path to personalised medicine in COPD. Thorax. 2014;69: 857–64.
18. Yang I, Clarke M, Eha S, Fong K. Inhaled corticosteroids for stable chronic obstructive pulmonary disease. Cochrane Database Syst Rev. 2012.
19. Han MK, Agusti A, Calverley PM et al. Chronic Obstructive Pulmonary Disease Phenotypes. Am J Respir Crit Care Med. 2010;182:598–604.
20. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. Lancet. 2008;372:1107–19.
21. Wardlaw AJ, Silverman M, Siva R, Pavord ID, Green R. Multi-dimensional phenotyping: Towards a new taxonomy for airway disease. Clin Exp Allergy. 2005; 35:1254–62.
22. Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining. 1st ed. Pearson Addison Wesley; 2006.
23. Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis. 5th ed. Wiley; 2011.
24. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Appl. Stat. 1979;28:100-8.
25. Calinski T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3:1–27.
26. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. 1963;58:236–44.
27. Wolpert DH, Macready WG. No Free Lunch Theorems for Optimisation. IEEE Trans Evol Comput. 1997;1:67–82.
28. Wolpert DH. Coevolutionary Free Lunches. IEEE Trans Evol Comput. 2005; 9:721–35.
29. Burgel PR, Paillasseur JL, Caillaud D et al. Clinical COPD phenotypes: A novel approach using principal component and cluster analyses. Eur Respir J. 2010;36: 531–9.
30. Garcia-Aymerich J, Gómez FP, Benet M et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. Thorax. 2011;66:430–37.
31. Bafadhel M, McKenna S, Terry S et al. Acute exacerbations of chronic obstructive pulmonary disease: Identification of biologic clusters and their biomarkers. Am J Respir Crit Care Med. 2011;184:662–71.
32. Burgel P-R, Paillasseur J-L, Peene B et al. Two Distinct Chronic Obstructive Pulmonary Disease (COPD) Phenotypes Are Associated with High Risk of Mortality. PLoS One. 2012;7:e51048.
33. Castaldi PJ, Dy J, Ross J et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. Thorax. 2014;69:415–22.
34. Rennard SI, Locantore N, Delafont B et al. Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. Ann Am Thorac Soc. 2015;12:303–12.
35. Esteban C, Arostegui I, Aburto M et al. Chronic obstructive pulmonary disease subtypes. transitions over time. PLoS One. 2016;11:1–16.
36. Chang Y, Glass K, Liu Y-Y et al. COPD subtypes identified by network-based clustering of blood gene expression. Genomics. 2016;107:51–8.
37. Burgel P-R, Paillasseur J-L, Janssens W et al. A simple algorithm for the identification of clinical COPD phenotypes. Eur Respir J. 2017;50:1701034.
38. Zarei S, Mirtar A, Morrow JD, Castaldi PJ, Belloni P, Hersh CP. Subtyping Chronic Obstructive Pulmonary Disease Using Peripheral Blood Proteomics. Chronic Obstr Pulm Dis. 2017;4:97–108.
39. Burgel PRG, Roche N, Paillasseur JL et al. Clinical COPD phenotypes identified by cluster analysis: Validation with mortality. Eur Respir J. 2012;40:495–6.
40. Dolnicar S. A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation. Australas. J Mark Res. 2003;2002:2–4.
41. Formann AK. Die Latent-Class-Analyse : Einführung in Theorie und Anwendung. Beltz; 1984.
42. Pagès J. Analyse factorielle de données mixtes. Rev Stat Appliquée. 2004;52:93–111.
43. Hopkins B. A New Method for determining the Type of Distribution of Plant Individuals. Ann Bot. 1954;112:213–27.
44. Fernández Pierna JA, Massart DL. Improved algorithm for clustering tendency. Anal Chim Acta. 2000;408:13–20.
45. Castaldi PJ, Benet M, Petersen H et al. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. Thorax. 2017;72:998–1006.
46. Miravitlles M, Soriano JB, García-Río F et al. Prevalence of COPD in Spain: Impact of undiagnosed COPD on quality of life and daily life activities. Thorax. 2009;64: 863–68.
47. Hvidsten SC, Storesund L, Wentzel-Larsen T, Gulsvik A, Lehmann S. Prevalence and predictors of undiagnosed chronic obstructive pulmonary disease in a Norwegian adult general population. Clin Respir J. 2010;4:13–21.
48. Gibson PG, McDonald VM. Asthma-COPD overlap 2015: Now we are six. Thorax. 2015;70:683–91.
49. Ghebre MA, Bafadhel M, Desai D et al. Biological clustering supports both "Dutch" and "British" hypotheses of asthma and chronic obstructive pulmonary disease. J Allergy Clin Immunol. 2015;135:63–72.
50. Ghebre MA, Pang PH, Diver S et al. Biological exacerbation clusters demonstrate asthma and chronic obstructive pulmonary disease overlap with distinct mediator and microbiome profiles. J Allergy Clin Immunol. 2018;141: 2027–36.
51. Rennard SI, Dale DC, Donohue JF et al. CXCR2 antagonist MK-7123 a phase 2 proof-of-concept trial for chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2015;191:1001–11.
52. Barnes P. Mediators of chronic obstructive pulmonary disease. Pharmacol Rev. 2004;56:515–48.
53. Stockley RA. Neutrophils and Protease / Antiprotease Imbalance. Am J Respir Crit. Care Med. 1999;160:S49-S52.
54. Wilson R, Sethi S, Anzueto A, Miravitlles M. Antibiotics for treatment and prevention of exacerbations of chronic obstructive pulmonary disease. J Infect. 2013;67:497–515.
55. Mohan A, Chandra S, Agarwal D et al. Prevalence of viral infection detected by PCR and RT-PCR in patients with acute exacerbation of COPD: A systematic review. Respirology. 2010;15:536–42.
56. Seemungal T, Harper-owen R, Bhowmik A et al. Respiratory Viruses, Symptoms, and Inflammatory Markers in Acute Exacerbations and Stable Chronic. Am J Respir Crit Care Med. 2001;164:1618–23.

57. Hospers JJ, Schouten JP, Weiss ST, Rijcken B, Postma DS. Asthma attacks with eosinophilia predict mortality from chronic obstructive pulmonary disease in a general population sample. Am J Respir Crit Care Med. 1999;160:1869–74.
58. Vedel-Krogh S, Nielsen SF, Lange P, Vestbo J, Nordestgaard BG. Blood Eosinophils and Exacerbations in Chronic Obstructive. Am J Respir Crit Care Med. 2016;193: 965–74.
59. Bafadhel M, Pavord ID, Russell REK. Eosinophils in COPD: just another biomarker? Lancet Respir Med. 2017;5:747–59.
60. Brightling CE, Bleecker ER, Panettieri Jr. RA et al. Benralizumab for chronic obstructive pulmonary disease and sputum eosinophilia: a randomised, double-blind, placebo-controlled, phase 2a study. Lancet Respir Med. 2014;2:891–901.
61. Dasgupta A, Kjarsgaard M, Capaldi D et al. A pilot randomised clinical trial of mepolizumab in COPD with eosinophilic bronchitis. Eur Respir J. 2017;49.
62. Pizzichini E, Pizzichini MMM, Gibson P et al. Sputum eosinophilia predicts benefit from prednisone in smokers with chronic obstructive bronchitis. Am J Respir Crit Care Med. 1998;158:1511–7.
63. Leigh R, Pizzichini MMM, Morris MM, Maltais F, Hargreave FE, Pizzichini E. Stable COPD: Predicting benefit from high-dose inhaled corticosteroid treatment. Eur Respir J. 2006;27:964–71.
64. Bafadhel M, Davies L, Calverley PMA, Aaron SD, Brightling CE, Pavord ID. Blood eosinophil guided prednisolone therapy for exacerbations of COPD: A further analysis. Eur Respir J. 2014;44:789–91.
65. Singh D, Kolsum U, Brightling CE, Locantore N, Agusti A, Tal-Singer R. Eosinophilic inflammation in COPD: prevalence and clinical characteristics. Eur Respir J. 2014;44:1697–1700.
66. Bafadhel M, Peterson S, De Blas MA et al. Predictors of exacerbation risk and response to budesonide in patients with chronic obstructive pulmonary disease: a post-hoc analysis of three randomised trials. Lancet Respir Med. 2018;6:117–26.
67. Regan EA, Hokanson JE, Murphy JR et al. Genetic Epidemiology of COPD (COPDGene) Study Design. Epidemiology. 2011;7:1–10.
68. Vestbo J, Anderson W, Coxson HO et al. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). Eur Respir J. 2008;31:869–73.

Usefulness of Biological Clustering Patterns in Chronic Obstructive Pulmonary Disease

Andreas Halner, BA and Mona Bafadhel, PhD, FRCP

Respiratory Medicine Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK

APPENDIX 1. Detailed summary of COPD phenotyping and endotyping studies

Study	Burgel et al. (2010) [29]	Garcia-Aymerich et al. (2011)[30]
Study hypotheses/ research questions	COPD subjects can be grouped into clinical phenotypes.	COPD includes various clinically relevant subtypes.
Size and characteristics of patient cohort	n = 322 Multi-centre study. Recruited patients were in a stable condition (defined as having no history of exacerbation requiring medication during the previous 4 weeks) and had a COPD diagnosis based on post-bronchodilator FEV_1/FVC < 0.7. Patients from all four GOLD severity categories were included. All patients in the cohort were positive for smoking history. Patients with a main diagnosis of asthma, bronchiectasis or other significant respiratory condition were excluded.	n = 342 Multi-centre study. Recruited patients had been hospitalised for a first time as a result of a COPD exacerbation. Measurements of variables were performed three months after discharge, at which point patients were clinically stable and COPD diagnosis was confirmed as post-bronchodilator FEV_1/FVC < 0.7.
Variables selected for dimension reduction and/or cluster analysis	A total of 8 variables were chosen for their relevance to pulmonary and extrapulmonary features of COPD: patient age, tobacco-smoking, FEV_1, number of exacerbations per patient per year, nutritional status, dyspnoea, health status and depressive symptoms.	A total of 536 variables, of which up to 150 variables were chosen for CA, were obtained relating to patient symptoms and quality of life, lung function tests, exercise capacity, patient nutritional status, biomarkers of systemic and bronchial inflammation, sputum microbiology, and imaging including CT of the thorax and echocardiography.
Method of analysis	PCA was performed to generate new independent variables – components – for CA. Ward's method of clustering was performed on 3 principal components since these components explained most of the variation in the data. Number of clusters was determined using the pseudo F and pseudo t^2 statistic.	K-means clustering was performed on the selected variables. Number of clusters was determined using the pseudo F statistic.
Cluster characteristics as described by author	Four clusters of patients were identified: 1: Younger patients with severe to very severe airflow limitation and respiratory symptoms but few comorbidities. 2: Older patients with mild respiratory disease and mild age-related comorbidities. 3: Younger patients with moderate-to-severe airflow limitation but milder symptoms and few comorbidities. 4: Older patients with moderate-to-severe airflow limitation, severe symptoms and significant cardiovascular comorbidity.	Three clusters of patients were identified: 1: Severe airflow limitation and very poor performance in other respiratory function tests. 2: Milder airflow limitation. 3: Milder airflow limitation with high proportion of comorbidity (obesity, cardiovascular disease, diabetes, systemic inflammation). Age of patients did not differ significantly between clusters.
Outcome measure for cluster validation and comparison	Longitudinal follow-up to determine all-cause mortality. Group 1 had the highest mortality rate[39].	Longitudinal follow-up to determine COPD-specific hospitalisation rate and all-cause mortality. Group 1 had the highest hospitalisation rate and all-cause mortality.

Study	Bafadhel et al. (2011) [31]	Burgel et al. (2012) [32]
Study hypotheses/ research questions	There are definable biological COPD clusters, which can be identified using biomarkers.	Subgroups of COPD patients differ in mortality.
Size and characteristics of patient cohort	n = 145 Single-centre study. Out of 145 patients entered into the study, 182 exacerbations were captured from 86 patients, and the desired variables were measured in 75 of these patients. All patients had post-bronchodilator FEV_1/FVC < 0.7 and COPD diagnosis. Patients from all four GOLD severity categories were included. Patients were all aged 40 years or older. Asthma or other lung disease apart from COPD was an exclusion criterion.	n = 527 Multi-centre study. Recruited patients had a COPD diagnosis based on post-bronchodilator FEV_1/FVC < 0.7. All patients were clinically stable. Patients from all four GOLD severity categories were included.
Variables selected for dimension reduction and/or cluster analysis	17 sputum biomarkers were considered for CA. Measurements were made for patients in the stable state and during exacerbations during the course of one year. Measurements of biomarkers at exacerbation were performed if the patients had not received prior oral corticosteroids or antibiotics.	7 continuous variables (patient age, BMI, FEV_1, dyspnoea, quality of life scale, thoracic gas volume and diffusing capacity) and numerous categorical variables (relating to comorbidities and imaging data e.g. CT analysis for emphysema).
Method of analysis	Factor analysis (a technique similar to PCA) was performed on the biomarkers, and 3 factors were selected for subsequent analysis since they accounted for the majority of the variation in the data. For each factor, the biomarker with the highest loading was used for CA. Ward's method was used to generate a dendrogram for visual inspection to select an appropriate number of clusters (4 was the value chosen). K-means clustering, pre-specified to identify 4 biological clusters, was then applied to the highest loading biomarkers. Receiver Operating Characteristic (ROC) curves were used to determine suitable biomarkers for identification of clinical phenotypes.	PCA and MCA were separately performed on continuous variables and on categorical variables, respectively. 2 principal components and 14 MCA axes were retained for Ward's clustering. Visual assessment of dendrogram was used to decide upon a suitable number of clusters.
Cluster characteristics as described by author	4 biological exacerbation clusters were identified: Cluster 1: A proinflammatory profile. Cluster 2: High levels of type 2 mediators. Cluster 3: High levels of type 1 mediators. Cluster 4: Low sputum mediator concentrations cluster which exhibited few changes in inflammatory profile. Clinically, clusters 1, 2 and 3 were bacteria-predominant, eosinophil-predominant and virus-predominant, respectively. Bacterial and eosinophilic clinical exacerbation phenotypes could be predicted from stable state.	Three clusters of patients were identified: 1: Airflow limitation and other respiratory disease features were mild to moderate, with few comorbidities. 2: Airflow limitation and other respiratory disease features were severe. Variable comorbidities - osteoporosis and muscle weakness common but cardiovascular comorbidities rare. 3: Airflow limitation and other respiratory disease features moderate to severe. Older patients. Obesity, diabetes and cardiovascular comorbidities common.
Outcome measure for cluster validation and comparison	Sensitive and specific biomarkers for clusters 1-3 were identified. Sputum IL1-β, percentage peripheral eosinophils and serum CXCL10 best identified the bacteria-, eosinophil- and virus-predominant subgroups, respectively. Validation of these biomarkers was performed in an independent cohort of 89 patients.	Longitudinal follow-up to determine all-cause mortality. Group 2 and 3 patients were at a significantly higher risk of mortality than group 1 patients.

(Continues on next page)

Appendix 1. Detailed summary of COPD phenotyping and endotyping studies *(Continued)*

Study	Castaldi et al. (2014)[33]	Rennard et al. (2015)[34]
Study hypotheses/ research questions	Distinct subtypes of pulmonary damage occur in smokers and these subtypes are strongly associated with relevant clinical outcome measures and COPD-associated genetic variants.	Clinically relevant subgroups of COPD exist and exhibit differences in relevant clinical outcomes when evaluated longitudinally.
Size and characteristics of patient cohort	n = 10192 Multi-centre study. Patients were examined at stable state (at least one month after last exacerbation). Patients from all four GOLD categories were included. Patients with a respiratory disease diagnosis other than COPD, asthma or emphysema were excluded. All recruited patients were between the ages of 45 and 80 and have a smoking history[67].	n = 2164 Multi-centre study. All recruited patients had a COPD diagnosis. Patients from GOLD categories 2 to 4 were included. Patients were all exacerbation-free for at least 4 weeks before inclusion in the study[68]. Patients with a respiratory disease diagnosis other COPD and severe alpha1-antitrypsin deficiency were excluded[68]. All recruited patients were between the ages of 40 and 75 and have a smoking history[68].
Variables selected for dimension reduction and/or cluster analysis	Different variables were used for different cluster models. The variables chosen for the final model were FEV_1, airway wall thickness and measures of emphysema from CT imaging.	41 variables relating to clinical, physiologic, imaging and biomarker parameters.
Method of analysis	Half of the patients constituted a training set to be included for CA, while the other half constituted a validation set to test the clusters. Various approaches were used to select variables for K-means clustering. The resulting clusters from these different models, as well as the different options for a suitable number of clusters, were compared against each other using normalised mutual information (NMI), a cluster stability measure. The different models were also compared for discriminatory power between relevant clinical outcomes. The best model consisted of 4 clusters, using only those variables which were uncorrelated (Pearson's correlation < 0.7).	Factor analysis was performed on the 41 variables. 13 factors were selected since they accounted for most of the variability in the data; variables with the highest loading for the 13 factors were chosen for CA. A random forest-based clustering approach was used and number of clusters chosen based on silhouette width and clinical relevance of the clusters.
Cluster characteristics as described by author	4 subgroups of patients were identified, all of which were reproduced in the validation group of patients: 1: Smoking-resistant patients with few symptoms of airways disease. 2: Patients with mild upper zone -predominant emphysema and airflow obstruction. 3: Patients with airway-predominant disease. 4: Patients with severe airway obstruction and emphysema.	5 subgroups of patients were identified: A: Patients with mild airways disease. B: Patients with intermediate values for health status and emphysema but low levels of inflammatory markers. C: Systemic inflammation, multiple comorbidities. D: Low FEV_1 and severe emphysema. E: Intermediate values for most variables.
Outcome measure for cluster validation and comparison	Clinical measures (including dyspnoea, exacerbation rates in previous year, hospitalisation rates in previous year) and COPD-associated gene variants. Subgroup 2, 3 and 4 had greater disease severity (subgroup 4 the worst) than subgroup 1 in terms of the validation clinical measures. After adjustment for the GOLD categories the patients in each subgroup were in, the association with clinical outcome measures remained significant. There was a strong association between subgroup 2 and the single nucleotide polymorphism (SNP) rs1980057 near the gene HHIP.	Longitudinal follow-up to determine all-cause mortality and time-to-first exacerbation. Subgroups C and D had the highest mortality and D had the shortest time-to-first exacerbation.

(Continues on next page)

Appendix 1. Detailed summary of COPD phenotyping and endotyping studies *(Continued)*

Study	Esteban et al. (2016)[35]	Chang et al. (2016[36]
Study hypotheses/ research question	Assess the stability of cluster-based COPD subgroups in patients with stable disease over 1 year.	Systemic inflammatory signals in peripheral blood gene expression data can identify clinically important COPD-related disease subtypes.
Size and characteristics of patient cohort	n = 543 Multicentre study. All patients had a previous COPD diagnosis for at least six months and were free from an exacerbation for at least 6 weeks prior to enrolling. For all patients, $FEV_1/FVC < 0.7$ and FEV_1/FEV_1 predicted < 0.8. Patients with an asthma diagnosis were excluded. One year after inclusion in the study, survivors were followed up.	n = 364 Multi-centre study. Data were captured at stable state. 229 former smokers (of whom 141 met the spirometric $FEV_1/FVC < 0.7$ criterion for COPD diagnosis) from the ECLIPSE cohort constituted the training set. Clusters obtained based on the gene expression data of these patients were then validated in a separate cohort of 135 smokers (of whom 76 met the spirometric $FEV_1/FVC < 0.7$ criterion for COPD diagnosis) from the COPDGene study.
Variables selected for dimension reduction and/or cluster analysis	A number of sociodemographic and clinical variables were included: age, BMI, number of previous hospitalisations, FEV_1%, hand strength, walking test, physical activity, dyspnoea, Charlson comorbidity index scores and occurrence of various comorbidities.	Gene expression patterns as determined from 1812 probesets associated with FEV_1 and FEV_1/FVC in the ECLIPSE study were used as input for cluster analysis.
Method of analysis	MCA was performed on the variables. 4 MCA factors were selected for hierarchical clustering. An optimal number of clusters was decided upon based on minimum loss of inertia. The aforementioned analyses were performed for patients at first inclusion in the study and then again after a year for surviving patients.	Network-based stratification (NBS) was used to achieve clustering of gene expression data. This was compared against a non-network based method (non-negative matrix factorisation); the former displayed superior performance to the latter as determined by stability indices associated with different numbers of latent factors. After clustering, subtype-specific gene expression for each cluster was analysed for gene ontology enrichment of known biological processes. The cluster model developed from gene expression of ECLIPSE cohort patients was used to generate clusters from the COPDGene cohort.
Cluster characteristics as described by author	4 clusters were identified: Cluster A: Less dyspnoea, better health-related quality of life and lower Charlson comorbidity scores. Cluster B: Intermediate between A and C. Cluster C: Most severe dyspnoea, and poorer pulmonary function and quality of life. Cluster D: Higher rates of hospitalization during the previous year; higher comorbidity scores. Whereas clusters A, B, and C, had marked respiratory profiles with a continuum in severity of several variables, cluster D was associated with a more systemic profile with intermediate respiratory disease severity.	4 clusters were identified in the ECLIPSE cohort via NBS and could be reproduced in the COPDGene cohort: Cluster 1: Enrichment for wound healing and inflammatory processes. Cluster 2: Enrichment for cytoskeletal and actin filament organization. Cluster 3: Enrichment for protein catabolism and ubiquitination. Cluster 4: Lymphocyte activation and protein synthesis.
Outcome measure for cluster validation and comparison	Clusters at baseline were compared with clusters after 1 year to assess stability. Clusters remained stable after 1 year, with only 28% patients migrating between the clusters. In addition, clusters were compared in terms of clinical outcomes such as mortality. Cluster D had the highest mortality after 1 year of follow-up, followed by cluster C, B and A respectively.	The gene-expression-based clusters were compared in terms of clinical characteristics and peripheral blood cell counts: Cluster 1: Most severe lung function impairment, respiratory symptoms, and emphysema. Higher levels of neutrophils than the other clusters. Cluster 2: Intermediate levels of lung function impairment, emphysema, and respiratory symptoms. Higher levels of eosinophils than the other clusters. Cluster 3 and cluster 4: Relatively preserved lung function. Cluster 3 had more emphysema, more respiratory symptoms, and a higher percentage of women than cluster 4. Cluster 3 and cluster 4 had higher levels of lymphocytes than the other clusters. Cell count differentials alone were not enough to predict COPD cluster membership. Smoking exposure did not differ significantly between the groups, suggesting that biological variability rather than cumulative smoke exposure, is more likely to explain the cluster patterns.

(Continues on next page)

Appendix 1. Detailed summary of COPD phenotyping and endotyping studies *(Continued)*

Study	Burgel et al. (2017)[37]	Zarei et al. (2017) [38]
Study hypotheses/ research questions	An algorithm can be developed for allocating COPD patients into CA-derived clinical phenotypes.	Peripheral blood proteomic data can be used to find subtypes of COPD within clinically similar individuals at stable state.
Size and characteristics of patient cohort	n = 6060 Multi-centre study. All recruited patients had a COPD diagnosis. Patients from all four GOLD categories were included. Patients were recruited in a stable state or at the time of a hospitalisation due to an exacerbation.	n = 396 Multi-centre study. All recruited patients were former smokers with at least 10 pack years but abstinence from smoking for at least 12 months before study. All patients had moderate to severe COPD (post-bronchodilator $FEV_1/FVC < 0.7$; FEV_1/FEV_1 predicted < 70% predicted; Diffusing capacity of the lung for carbon dioxide capacity (DLCO) < 0.7) and were at stable state. All patients had emphysema based on visual examination of CT scans.
Variables selected for dimension reduction and/or cluster analysis	Clinical variables including age, BMI, FEV_1, dyspnoea, number of exacerbations in previous 12 months, presence of comorbidities (cardiovascular and diabetes).	87 protein biomarkers measured from peripheral blood were used as input for CA.
Method of analysis	Factor analysis for mixed data (FAMD) was performed on the variables of 2409 patients and the highest loading variables for the selected factors were used as input for Ward's method of clustering. Classification and regression trees (CARTs) were then trained to develop an algorithm for allocating patients into the clusters obtained from Ward's method in the group of 2409 patients. The algorithm was then tested on a separate cohort of 3651 patients.	Agglomerative McQuitty hierarchical clustering was performed on the biomarker dataset. The optimal number of clusters was determined through the R package NbClust which takes into account a variety of indices for this purpose. After clustering, enrichment analysis was performed for the biomarkers which had different mean values among the clusters in order to identify the molecular pathways associated with these biomarkers.
Cluster characteristics as described by author	Five subgroups were identified: I: Severe respiratory disease, older age patients, high prevalence of comorbidities. II: Moderate-to-severe respiratory disease, younger patients, few comorbidities. III: Older patients, high prevalence of comorbidities. IV: Very severe respiratory disease, younger patients, few comorbidities. V: Mild respiratory disease, younger patients, few comorbidities.	3 stable state biological clusters were identified. The biomarkers which distinguished cluster 3 from cluster 1 and cluster 2 mapped to platelet alpha granule and cell chemotaxis pathways.
Outcome measure for cluster validation and comparison	Longitudinal follow-up to determine all-cause mortality. Subgroups I and IV have highest mortality rate. The CART-developed algorithm successfully assigned patients in the validation cohort to five classes which corresponded to the clusters obtained in the training cohort in terms of the relative clinical characteristics and mortality rates between the classes.	The clusters were compared in terms of clinical and physiological characteristics. Compared to cluster 1 and cluster 2, cluster 3 had less emphysema on quantitative analysis of chest CT scans and worse disease-related quality of life based on the St. George's Respiratory Questionnaire.

BMI: body mass index; CA: cluster analysis; COPD: chronic obstructive pulmonary disease; CT: computed tomography; ECLIPSE: Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points; FEV_1: forced expiratory volume in one second; FVC: forced vital capacity; GOLD: Global Initiative for Chronic Obstructive Lung Disease; IL: interleukin; MCA: multiple correspondence analysis; PCA: principal component analysis.

APPENDIX 2. Detailed examples of cluster analysis studies including both asthma and COPD

Study	Ghebre et al. (2015)[49]	Ghebre et al. (2018)[50]
Study hypotheses/ research questions	Determine the extent to which COPD and asthma at stable state represent distinct or overlapping conditions in terms of sputum cellular and mediator profiles.	Investigate the sputum cellular, mediator, and microbiome profiles of asthma and COPD exacerbations via a CA-based approach.
Size and characteristics of patient cohort	n = 385 Single-centre study. 161 patients (86 with severe asthma and 75 with moderate-to-severe COPD) were recruited to constitute the training set for CA. Patients were examined at stable state (defined here as at least 8 weeks free from an exacerbation). 224 patients (166 with severe asthma and 58 with COPD) were used for validation.	n = 105 Single-centre study. 32 asthmatic patients and 73 patients with COPD were recruited. Patients were assessed at exacerbation.
Variables selected for dimension reduction and/or cluster analysis	21 sputum mediators were analysed in the training study and 14 in the validation study.	19 sputum mediators were analysed.
Method of analysis	Factor analysis was performed and 4 factors selected based on scree plots and having an eigenvalue > 1. K-means analysis was then performed on the factor scores. The optimal number of clusters was selected based on a scree plot and the biological/clinical interpretability of the resulting clusters.	Factor analysis was performed and 3 factors selected based on scree plots and having an eigenvalue > 1. K-means analysis was then performed on the factor scores. The optimal number of clusters was selected based on a scree plot.
Cluster characteristics as described by author	3 stable state biological clusters were identified and reproduced in the validation group of patients: Cluster 1:Aasthma-predominant, eosinophilic, high TH2 cytokines. Cluster 2: Asthma and COPD overlap, neutrophilic; this cluster was associated with the highest overall inflammatory cell count. Cluster 3: COPD-predominant, mixed eosinophilic and neutrophilic.	3 exacerbation state biological clusters were identified: Cluster 1: COPD predominant. 27 patients with COPD and 7 asthmatic patients exhibited increased blood and sputum neutrophil counts, proinflammatory, and proportions of the bacterial phylum Proteobacteria. Cluster 2: 10 asthmatic patients and 17 patients with COPD with increased blood and sputum eosinophil counts, type 2 mediators, and proportions of the bacterial phylum Bacteroidetes. Cluster 3: 15 asthmatic patients and 29 patients with COPD with increased type 1 mediators and proportions of the phyla Actinobacteria and Firmicutes.
Outcome measure for cluster validation and comparison	The clusters were compared in terms of clinical, physiological and biological characteristics. Linear discriminant analysis was performed on the inflammatory mediators and a model developed to group patients into the training set clusters. This model was then used to categorise patients of the validation set into inflammatory mediator-based clusters. Patients in the asthma-COPD overlap cluster had higher symptoms of cough than patients of the other two clusters.	The clusters were compared in terms of clinical, physiological and biological characteristics. Linear discriminant analysis was performed on the inflammatory mediators to develop a model to separate patients out by cluster. However, this model was not validated in a separate patient test set.

CA: cluster analysis; COPD: chronic obstructive pulmonary disease.