REVIEW

Vasilis Nikolaou et al

COPD phenotypes and machine learning cluster analysis: A systematic review and future research agenda

Vasilis Nikolaou[1], Sebastiano Massaro[1,2], Masoud Fakhimi[1], Lampros Stergioulas[1], David Price[3]

[1] Surrey Business School, University of Surrey, Guildford GU2 7HX, United Kingdom

[2] The Organizational Neuroscience Laboratory, London WC1N 3AX, United Kingdom

[3] Observational and Pragmatic Research Institute, Singapore, Singapore; Centre of Academic Primary Care, Division of Applied Health Sciences, University of Aberdeen, Aberdeen, United Kingdom

Correspondence: Vasilis Nikolaou

University of Surrey, Surrey Business School, Alexander Fleming Rd, Guildford GU2 7XH, United Kingdom

Tel : + 44 7799 363802

Email : v.nikolaou@surrey.ac.uk

Abstract: Chronic Obstructive Pulmonary Disease (COPD) is a highly heterogeneous condition projected to become the third leading cause of death worldwide by 2030. To better characterize this condition, clinicians have classified patients sharing certain symptomatic characteristics, such as symptom intensity and history of exacerbations, into distinct phenotypes. In recent years, the growing use of machine learning algorithms, and cluster analysis in particular, has promised to advance this classification through the integration of additional patient characteristics, including comorbidities, biomarkers, and genomic information. This combination would allow researchers to more reliably identify new COPD phenotypes, as well as better characterize existing ones, with the aim of improving diagnosis and developing novel treatments. Here, we systematically review the last decade of research progress, which uses cluster analysis to identify COPD phenotypes. Collectively, we provide a systematized account of the extant evidence, describe the strengths and weaknesses of the main methods used, identify gaps in the literature, and suggest recommendations for future research.
Keywords: chronic respiratory disease, subtypes, statistical analysis

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a group of lung diseases, such as emphysema, chronic bronchitis, and asthma, that cause breathing difficulties due to inflammation of the lungs and narrowing of the airways. Typical symptoms of COPD include breathlessness, a persistent cough with phlegm, frequent chest infections, and wheezing. Its main causes are smoking, which accounts for almost 90% of cases, occupational exposure to dust and fumes, and air pollution [1]. COPD represents one of the most common respiratory diseases, and it is projected to become the third leading cause of death worldwide by 2030 [2], principally because of difficulties in early, accurate diagnosis.

To better characterize COPD and improve diagnosis, the extant research has identified different patient phenotypes (i.e., the common clinical characteristics shared by patients affected by COPD). These phenotypes are usually assessed through clinical examinations, generally following the recommendations provided by the Global Obstructive Lung Disease initiative (GOLD) [3]. Specifically, GOLD classifies COPD patients into four phenotype-like categories according to a 2x2 matrix structured along the dimensions of symptoms and history of exacerbations (Table 1).

[Table 1 about here]

Whilst beneficial in guiding clinical practice, this and other forms of COPD classification are often in need of stronger statistical support with respect to their predictive ability regarding clinically meaningful outcomes, such as mortality and response to treatment [4]. For instance, a large prospective study (n=12,108 patients) recently showed that COPD patients receiving maintenance therapy were similarly distributed across the four GOLD phenotypes when compared to patients who received a target treatment [5]. Likewise, the proportion of comorbidities and rate of exacerbations reported across the COPD groups were similar for both cohorts, suggesting a limited discriminatory ability of these phenotypes [5].

To address this issue, research has increasingly called for the integration of other determinants, such as physiological characteristics (e.g., age, body mass index, waist circumference) [6-16,18], comorbidities (e.g., diabetes, cardiovascular diseases) [6,8,10,13,16,17,19], pulmonary function tests [7,8,11-16,19],biomarkers [6,19], and genetic variants [7], as valuable information to facilitate a more comprehensive characterization of the distinctive biological nature of COPD phenotypes, thereby promising to improve their predictive ability for clinically relevant outcomes. In particular, with sustained progress in applying machine learning algorithms to medicine, research has recently begun to put forward the powerful method of clustering – a machine learning method, which allows researchers to find structures in the data so that the elements of the same cluster (i.e., a phenotype) are more similar to each other than to those from different clusters [20], with the aim of integrating patients' information and identifying patterns of association that can characterize COPD phenotypes more precisely.

Yet, at present, there is still little evidence-based information available that both systematizes current knowledge on cluster analysis for COPD phenotype characterization and pinpoints the core benefits and limitations of the different approaches. Here, we aim to tackle this gap by reviewing the last decade of research, which uses cluster analysis to identify clinically meaningful COPD phenotypes. In the following sections of this article we describe our search strategy, synthetize the characteristics of the articles retrieved (e.g., study design, population, phenotypes' features), and provide recommendations aimed at improving the use and performance of these methods in future research and clinical practice.

Search strategy and selection criteria

In keeping with PRISMA guidelines, we conducted our search through a systematic consultation of the Medline PubMed, Cochrane Library, Scopus, and Web of Science (Figure 1) databases.

[Figure 1 about here]

We also hand-searched the reference lists of the retrieved articles. Additionally, we searched articles in leading pulmonary and respiratory medicine scholarly outlets to specifically include journals such as The Lancet Respiratory Medicine and The American Journal of Respiratory and Critical Care Medicine, among others.

Briefly, we tailored the search to probe for overarching concepts and relations pertaining to the domains of machine learning and COPD phenotypes. Specifically, we searched for studies that used cluster analysis to identify COPD phenotypes by using the MeSH keywords “COPD”, “phenotypes”, “cluster analysis”, “clustering” and “machine learning” as well as their possible variants and combinations. Moreover, we aimed to search for articles in which the COPD phenotypes reported were validated by clinically meaningful outcomes, eg, mortality, exacerbations, and response to therapy. We also searched for ongoing registered studies relevant to our research question, including NOVELTY [21], SPIROMICS [22] and the BigCOPData [23] project, which, whilst informative to the overall picture, were not individually retained in our analysis because their final results have yet to be fully disclosed.

Our search resulted in 117 articles published mainly in English and covering the period between 2003 and 2019. After excluding duplicates, we screened 113 papers to select unambiguous publications of relevant research. Hereby, 65 articles were excluded because they were not relevant to COPD phenotypes and/or machine learning methods, while 34 studies were excluded because the COPD phenotypes reported had not been validated with clinically meaningful outcomes.

Fourteen studies that satisfied our inclusion/exclusion criteria were retained in this review. Next, we present the entire body of retrieved studies, focusing in particular on the population characteristics, study design, sample size, the derived COPD phenotypes, and the clinical outcomes against which the phenotypes were validated of the articles respecting our inclusion criteria (Table 2). Moreover, we highlighted important inputs that we appreciated from the studies excluded from our systematic analysis, as well as specific phenotypes observed in the Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE) [24] study.

[Table 2 about here]

Findings

Studies respecting inclusion criteria for review

Populations

The sample size varied considerably across studies, spanning from 65 [18] to 30,961 patients [6].The majority of the retrieved works involved multi-centre, observational cross-sectional cohorts across the world (e.g., Italy, France, Spain, Belgium, United Kingdom, Korea, Japan, New Zealand, China). Data were collected from university hospitals, tertiary care, and pulmonary rehabilitation settings. This variability may explain the high variation in sample sizes. For instance, the largest study [6] (ie, CALIBER) covered a longitudinal cohort for a period of 18 years. This cohort comprised the data of electronic health records from three UK national resources: the Clinical Practice Research Datalink (CPRD), the Hospital Episode Statistics (HES), and information on cause-specific mortality from the Office for National Statistics (ONS). The second largest study [7] was based on the Genetic Epidemiology of COPD (COPDGene) and aimed to investigate the genetic factors responsible for COPD development. Moreover, similar to CALIBER [6], Burgel et al [8] combined three national COPD cohorts from France and Belgium as well as one independent cohort from the COPD Cohorts Collaborative International Assessment (3CIA) initiative. Two other relatively large studies, each with over 1,000 patients, were carried out in Asia. One was based on the Korean COPD subgroup multi-centre cohort [9] and the other one [10] included out-patients of universities' pulmonary clinics and referral hospitals in 13 Asian cities.

Importantly, despite the diverse ethnic backgrounds of the populations of these studies, the identified COPD phenotypes were rather consistent across studies, including elements of asthma-COPD overlaps, comorbidities, and obesity, amongst others.

Clinical Outcomes

A core characteristic shared among the reviewed studies is that all COPD phenotypes were validated by clinically meaningful outcomes, such as exacerbations, health-related quality of life, mortality rate, and responses to therapy. These phenotypes were cross-validated in a large (n=2,746) three-year observational multi-centre international study – the Evaluation of COPD Longitudinally to Identify

Predictive Surrogate End-points (ECLIPSE) [24]. In this study, a cross-sectional analysis of the baseline data showed that patients with COPD had more frequent comorbidities, especially cardiovascular ones, when compared to controls [25]. It also showed that males with COPD were more susceptible to cardiovascular comorbidity than females; moreover, in Pikoula et al [6], patients with comorbid cardiovascular disease and diabetes were characterized by high hospital admission rates for acute exacerbations of COPD (AECOPD) and were reported as being more likely to die of cardiovascular disease.

Building on these results, subsequent works [26,27] identified phenotypes of patients with frequent (i.e., two or more per year) exacerbations as well as patients with a rapid decline in their lung function. The latter evidence [27] was further extended by a five-year longitudinal study that classified patients into three groups: fast decline, slow decline, and stable patterns [28]. The latter work showed that the only factor significantly associated with a fast decline of FEV1 (Forced Expiratory Volume in 1 second) was the severity of the emphysema. Moreover, 25% of the cohort was represented by the so-called “asthma-COPD overlap,” or ACO, in which patients are characterized by having more exacerbations and more frequent comorbidities than in other rapid-decline COPD types [29].

Features of COPD Phenotypes

We found substantial heterogeneity in both the numbers and features of phenotypes presented in the literature. The number of COPD phenotypes identified varied from two to five, the most frequently reported being three [10,11,13-15] and five [6,8,16,17,19].

Intriguingly, the features pertaining to the three most reported phenotypes varied across studies. For instance, phenotypes were characterized by patients having frequent exacerbations and a fast decline in lung function and in quality of life [10], but also by patients of a young age with fewer symptoms and exacerbations [11], or patients with severe respiratory disease but a low rate of comorbidities and older patients with a high rate of comorbidities (e.g., cardiovascular diseases and diabetes) but lower airway limitation and less obesity [12,13].

Two studies [14,15] reported similar phenotypes with respect to COPD severity. Peters et al [14] identified three phenotypes in which patients were characterized by moderate COPD and a low impact on overall health status, moderate COPD with a high impact on health status, or severe COPD with a moderate impact on health status. Similarly, the three phenotypes identified by Garcia-

Aymerich et al [15] were characterized by moderate, severe, and systemic COPD; the latter phenotype also had a high rate of cardiovascular comorbidities.

When four phenotypes were reported, they also differed in terms of the severity of symptoms. Specifically, Yoon et al [9] clustered patients both according to their COPD severity (ie, mild, moderate, severe) and by identifying the ACO phenotype. A related work [7] classified patients according to the severity of emphysema (i.e., minimal, moderate, severe). Moreover, two studies [12,13] emphasized the distinction of two key population groups: a younger group of patients with moderate to severe respiratory disease but few comorbidities, and an older group with mild to severe airflow limitations but a high rate of cardiovascular comorbidities.

In those articles that identified five phenotypes, the reported features were more homogeneous than those identified in studies reporting fewer phenotypes. For instance, almost every study reported similar comorbidities, namely cardiovascular and metabolic diseases (e.g., diabetes), obesity, and ACO, as possible confounding factors. In Burgel et al [8], the derived phenotypes confirmed other existing findings [12,13], suggesting the identification of an older group of patients with a high rate of cardiovascular comorbidities and diabetes but with less severe respiratory impairments. Similarly, Chen et al [16] acknowledged a group of young patients with mild airflow obstructions, few symptoms, and infrequent severe exacerbations vis-à-vis older patients with more symptoms, frequent severe exacerbations, and a high mortality rate.

Overall, the diversity of phenotypes and populations presented in the current literature should not be surprising. Indeed, as we explain in the following, this scenario is largely due to an overarching limited reliance on statistical support in validating COPD with clinically meaningful outputs. Confirming our argument, for instance, a large study [30] carried out across ten independent cohorts from different populations in North America and Europe clearly showed that when identical methods were implemented for 17,146 individuals with COPD using common COPD-related characteristics, the reproducibility of COPD phenotypes across studies was rather modest.

Studies excluded from the systematic analysis

Ninety-nine studies were excluded either because a) they were irrelevant to COPD phenotypes or machine learning methods under study or b) the reported COPD phenotypes were not validated against clinical meaningful outcomes (Table 3).

[Table 3 about here]

Twenty one of those studies identified between two [31] and nine [32] phenotypes; however the number of phenotypes most frequently reported were either three [33, 34, 35, 36], four [37, 38, 39, 40, 41, 42, 43, 44] or five [45, 46, 47, 48, 49, 50, 51]. The works were predominantly observational – 12 were cross-sectional [31, 33, 36, 38, 39, 40, 41, 42, 47, 48, 49, 51], six prospective [34, 43, 44, 45, 46, 50], two retrospective [32, 37] and one randomised placebo controlled clinical trial [35]. Reported samples were comprised between 75 [36] and 3,144 [32] patients. In these studies, there was a remarkable heterogeneity among the reported phenotypes. For instance, when three phenotypes were reported, patients were characterized as either being young with few symptoms and mild airway limitation, or older and highly symptomatic with severe airway limitation or as a combination of both [34]. Moreover, de Torres et al. [34] showed that these phenotypes remained stable in most of the patients over a two years follow-up period.

In studies with four phenotypes patients were characterized by the severity of the disease, i.e., patients with mild to moderate disease, moderate to severe emphysema, mild to increased dyspnoea, low to high exacerbation risk or even an overlap of asthma and COPD [38, 39, 41]. In one of these studies, Bafadhel et al [43] classified patients into four biologic clusters: a) bacterial-predominant, b) viral-predominant, c) eosinophilic-predominant and d) patients with limited changes in their inflammatory profile.

In clusters of five phenotypes patients were characterized not only by the severity of the disease [45, 48] but also by the presence of comorbidities [46] as well the asthma and COPD overlap syndrome [47, 48, 49]. We also observed a reported distinction between female patients with high body mass index, asthma, COPD, and symptom scores but no inflammation, and male patients with asthma and COPD with high eosinophil counts and low use of oral corticosteroids [47]. Another salient difference was shown between younger-onset asthma patients with severe symptoms and elderly patients with high frequency of comorbidities and concomitant COPD [50].

A list of all potential phenotypes along with their groupings is displayed in the Appendix. Although this list is not exhaustive, it summarizes the most frequently reported phenotypes of the reviewed studies.

Methods

Study design

Generally speaking, the retrieved research based on observational studies [6-8] highlights the advantage of capturing large cohorts of patients with COPD as well as the opportunity to showcase "real-life" outputs from clinical practice. Moreover, and in contrast to controlled experiments such as clinical trials in which patients are selected homogeneously to satisfy certain inclusion and exclusion criteria, an observational study allows researchers to appreciate the patients' heterogeneity, which is a defining feature of COPD. Hence, the analysis of and outputs from such studies advance knowledge with respect to sample representativeness, covering actual COPD populations from different geographical settings.

On the other hand, the results coming from observational studies may lead to the emergence of unstable phenotypes, in turn making treatment decisions more complex. Similarly, because observational studies are generally carried out in university hospitals, tertiary care centres or rehabilitation settings, they tend to cover only severe COPD patients and may not be fully representative of the wider COPD population.

Validation

Across the reviewed studies, we acknowledge that the derived COPD phenotypes were often validated both internally (i.e., from the same population in terms of clinically meaningful outcomes such as exacerbations, mortality, and response to therapy) and externally on a different population (e.g., including the rapid lung function decline or the asthma-COPD overlap phenotype in the ECLIPSE cohort). This procedure offers strong reliability as it provides evidence for the generalizability and robustness of the results.

Data reduction and clustering

Most of all, from our analysis of the literature, we can appreciate the recurrent use of statistical techniques aiming to reduce the size of the data and group patients with similar characteristics into distinct clusters. These approaches have the immediate advantage of utilizing all available information, yet in practice they "operationalize" phenotypes as if they were mathematical constructs and as a result they may not always be closely relevant to the medical condition.

As such, issues such as the handling of missing data or the choice of variables feeding the analysis become paramount features to ensure the consistency of phenotype identification in progressing with COPD research. For instance, while the analysis of common features already offers a moderate concordance in determining COPD phenotypes [30], their robustness and reproducibility using an extended or diverse list of variables remains to be determined.

We argue that one of the first steps needed to overcome the issue of ensuring the reproducibility and alignment of COPD phenotypes is situated, at least to some extent, in the variety of statistical methods used to derive them (Table 4).

[Table 4 about here]

Most of the reviewed literature used data reduction methods to select the variables to include in the cluster analysis [6-8,10-13,16]. These methods vary from Principal Components Analysis (PCA) [52] to Multiple Correspondence Analysis (MCA) [53] – a method similar to PCA yet using categorical data – and factor analysis. PCA, MCA, and factor analysis [54, 55] share the characteristic that they reduce data dimensionality to identify a small number of clinically relevant variables able to explain most of the variations occurring in COPD patients' data. Whilst these approaches are beneficial to summarize data with a few variables without losing information, the interpretation of the derived variables within a clinical context is rarely straightforward due to their intimate mathematical nature. Other studies [9,14,15,17,19] selected variables on either data availability and/or clinical expertise, i.e., by including a priori available variables deemed to be relevant to COPD alone. For instance, Chubachi et al [17] used only comorbidity data, while others used either a combination of lung function and demographic data (i.e., age, BMI, smoking status) [9,12,14,16,18] or a combination of lung function, demographic, comorbidity, and biomarker data [6,8,10,19]. Thus far, only a few articles combined all the above information with imaging and/or genetic data [7,11,13,15]. The variability in the choice of variables can thus lead to the unpredictability of the characteristics of the derived phenotypes.

Noticeably, seven works used hierarchical analysis [8,10,11-13,17,19], which is a method in which each cluster is part of a larger cluster and they are all connected to each other like a tree (or dendrogram), whereby the number of clusters is determined by visual inspection [56]. Four studies [7,9,15,18] used k-means clustering, a method that splits the data into mutually exclusive clusters and in which the number of clusters needs to be specified in advance. Finally, two studies [6,16] used a

combination of hierarchical and k-means clustering, and one [14] used a combination of hierarchical and discriminant analysis, a technique that discriminates the categories of a dependent variable (e.g., symptoms) and evaluates the accuracy of this classification.

Missing values

We note that regardless of the method used, an important aspect of these cluster analysis approaches is the handling of missing values. Indeed, most of the reviewed studies failed to address this issue. Research tended to use non-missing data to form COPD clusters without considering which phenotypes might have been formed if patients with missing data had been included in the analysis or if only a portion of them had been excluded. Only two studies [6,15] considered alternative methods for assessing the impact of excluding patients on the formation of COPD phenotypes. Pikoula et al [6] performed a sensitivity analysis by excluding all patients with a diagnostic code for asthma and identified four clusters. Notably, the atopic cluster did not present a strong enough discriminant ability to form a separate cluster. Thus, atopic patients were categorized as belonging to either the anxiety/depression or the not-comorbid phenotype. Garcia-Aymerich [15] instead considered the use of multiple imputation when implementing the cluster analysis [57], allowing simulated values to replace the missing ones and thereby enabling the use of data from all patients.

Discussion

There are several implications of clinical and medical relevance in using machine learning methods to extract data from different sources, such as radiology, imaging or genetics, to identify clinically relevant COPD phenotypes. In sum, these include a better understanding of the natural history of the disease, the opportunity to more accurately identify high risk patient profiles, the prospect of early diagnosis and target treatments specific to certain phenotypes - along with the limitation of potentially adverse effects of unnecessary treatments, and the ability to make better and more precise predictions of treatment outcomes, thereby improving the prognosis of the disease and optimizing the use of health care resources.

Building on the evidence emerging from this review, we can identify several recommendations for future research using cluster analysis to identify COPD phenotypes; these are summarized in Table 5.

These strategies include the use of large samples to make clinically meaningful associations and the handling of missing data to assess the robustness of the results.

[Table 5 about here]

Moving forward, in keeping with Bourbeau et al. [58], we suggest that regardless of the clustering method chosen, COPD-derived phenotypes should be validated both internally and externally. This aspect is central because clustering methods are data-driven techniques, thus the derived clusters might be subject to spurious groupings.

As such, best practices in deriving COPD phenotypes include the utilization of prospective longitudinal data, which allows the assessment of variability and stability of features over time, as well as the use of cohorts from different settings to obtain the full spectrum of COPD phenotypes. The former recommendation implies carrying out large observational longitudinal cohort studies with at least a 3-year follow-up, as currently seen in the CALIBER [6] and ECLIPSE [24] studies. The latter proposal suggests using cohorts from different populations and settings to fully capture the heterogeneity of COPD. In this respect, we also envision the benefit of analysing cohorts including genetic information, such as COPDGene [7] or the UK Biobank database [59]. The immediate advantages of using such databases will be the opportunity to analytically and jointly assess patients' clinical characteristics (eg, lung functionality), comorbidities, and biomarker data to strengthen the robustness of the COPD phenotypes as well as to better understand the underlying biological mechanisms of the condition.

Ensuring clarity in the choice of variables used for identifying COPD phenotypes is another crucial recommendation for research using cluster analysis. This selection should always be evidence-based through experts' opinions and/or published works to avoid choosing variables that might not be clinically relevant [58]. At the same time, we recognize that this approach may lead to previously unidentified patient characteristics being overlooked. Thus, we suggest that a reasonable compromise moving forward would be to use available evidence alongside clustering analysis. As such, the combination of hierarchical, k-means clustering, and clinical judgment appears to be the most suitable approach to specify the correct number of clusters leading to the identification of novel COPD phenotypes.

Conclusions

This article reviewed research published in the last decade on COPD phenotypes identified using cluster analysis and validated with clinically meaningful outcomes. To the best of our knowledge, this is one of the first works addressing such a systematization of the COPD literature. Moreover, it puts forward key recommendations to improve the study design, variables selection, external validation, and handling of missing data of prospective studies.

Finally, we believe that future research should be tasked with further investigating COPD phenotype(s) whose characteristics have not yet been fully explored. For instance, the “fast decliner” phenotype [10,26,27], characterized by young patients with COPD with a fast decline in their lung function, as well as the cardiovascular comorbidity [6,13,25] characterized by differences in age, sex and high rates of hospital admission for AECOPD represent promising issues which are still largely unaddressed. Whichever the phenotype, we are hopeful that the insights presented here will soon enable research to better characterize additional patient determinants of COPD phenotypes and explore their association with responses to therapy while possibly developing more targeted treatments.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. NHS inform on Chronic obstructive pulmonary disease. https://www.nhsinform.scot/illnesses-and-conditions/lungs-and-airways/copd/chronic-obstructive-pulmonary-disease#about-copd (accessed February 15, 2020)
2. World Health Organization on chronic respiratory diseases and COPD. https://www.who.int/respiratory/copd/en/ (accessed February 15, 2020)
3. Global Initiative for Chronic Obstructive Lung Disease. Pocket guide to COPD diagnosis,

management and prevention. 2019 Report. https://goldcopd.org/wp-content/uploads/2018/11/GOLD-2019-POCKET-GUIDE-FINAL_WMS.pdf (accessed February 15, 2020)

4. Burgel PR, Paillasseur JL, Roche N. Identification of clinical phenotypes using cluster analyses in COPD patients with multiple comorbidities. BioMed research international. 2014;2014.
5. Halpin DM, de Jong HJ, Carter V, Skinner D, Price D. Distribution, temporal stability and appropriateness of therapy of patients with COPD in the UK in relation to GOLD 2019. EClinicalMedicine. 2019 Sep 1;14:32-41.
6. Pikoula M, Quint JK, Nissen F, Hemingway H, Smeeth L, Denaxas S. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. BMC medical informatics and decision making. 2019 Dec;19(1):86.
7. 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 May 1;69(5):416-23.
8. Burgel PR, Paillasseur JL, Janssens W et al. A simple algorithm for the identification of clinical COPD phenotypes. European Respiratory Journal. 2017 Nov 1;50(5):1701034.
9. Yoon HY, Park SY, Lee CH et al. Prediction of first acute exacerbation using COPD subtypes identified by cluster analysis. International journal of chronic obstructive pulmonary disease. 2019;14:1389.
10. Kim WJ, Gupta V, Nishimura M et al. Identification of chronic obstructive pulmonary disease subgroups in 13 Asian cities. The International Journal of Tuberculosis and Lung Disease. 2018 Jul 1;22(7):820-6.

11. Kim S, Lim MN, Hong Y, Han SS, Lee SJ, Kim WJ. A cluster analysis of chronic obstructive pulmonary disease in dusty areas cohort identified three subgroups. BMC pulmonary medicine. 2017 Dec;17(1):209.
12. Burgel PR, Paillasseur JL, Caillaud D et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. European Respiratory Journal. 2010 Sep 1;36(3):531-9.
13. Burgel PR, Paillasseur JL, Peene B et al. Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality. PloS one. 2012;7(12).
14. Peters JB, Boer LM, Molema J, Heijdra YF, Prins JB, Vercoulen JH. Integral health status-based cluster analysis in moderate-severe copd patients identifies three clinical phenotypes: Relevant for treatment as usual and pulmonary rehabilitation. International journal of behavioral medicine. 2017 Aug 1;24(4):571-83.
15. Garcia-Aymerich J, Gómez FP, Benet M et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. Thorax. 2011 May 1;66(5):430-7.
16. Chen CZ, Wang LY, Ou CY, Lee CH, Lin CC, Hsiue TR. Using cluster analysis to identify phenotypes and validation of mortality in men with COPD. Lung. 2014 Dec 1;192(6):889-96.
17. Chubachi S, Sato M, Kameyama N et al. Identification of five clusters of comorbidities in a longitudinal Japanese chronic obstructive pulmonary disease cohort. Respiratory medicine. 2016 Aug 1;117:272-9.
18. Altenburg WA, de Greef MH, Ten Hacken NH, Wempe JB. A better response in exercise capacity after pulmonary rehabilitation in more severe COPD patients. Respiratory

medicine. 2012 May 1;106(5):694-700.

19. Fingleton J, Travers J, Williams M et al. Treatment responsiveness of phenotypes of symptomatic airways obstruction in adults. Journal of Allergy and Clinical Immunology. 2015 Sep 1;136(3):601-9.

20. Everitt B. (1974). Cluster Analysis Heinemann. London.

21. Reddel HK, de Verdier MG, Agustí A et al. Prospective observational study in patients with obstructive lung disease: NOVELTY design. ERJ open research. 2019 Feb 1;5(1):00036-2018.

22. Study of COPD Subgroups and Biomarkers (SPIROMICS). [ClinicalTrials.gov Identifier: NCT01969344]

23. Chart Review of Patients With COPD, Using Electronic Medical Records and Artificial Intelligence (BigCOPData) [ClinicalTrials.gov Identifier: NCT04206098]

24. Papi A, Magnoni MS, Muzzio CC, Benso G, Rizzi A. Phenomenology of COPD: interpreting phenotypes with the ECLIPSE study. Monaldi Archives for Chest Disease. 2016 Oct 14;83(1-2).

25. Agusti A, Calverley PM, Celli B et al. Characterisation of COPD heterogeneity in the ECLIPSE cohort. Respiratory research. 2010 Dec 1;11(1):122.

26. Hurst JR, Vestbo J, Anzueto A et al. Susceptibility to exacerbation in chronic obstructive pulmonary disease. New England Journal of Medicine. 2010 Sep 16;363(12):1128-38.

27. Vestbo J, Edwards LD, Scanlon PD et al. Changes in forced expiratory volume in 1 second over time in COPD. New England Journal of Medicine. 2011 Sep 29;365(13):1184-92.

28. Nishimura M, Makita H, Nagai K et al. Annual change in pulmonary function and clinical phenotype in chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine. 2012 Jan 1;185(1):44-52.

29. Donohue JF, Herje N, Crater G, Rickard K. Characterization of airway inflammation in patients with COPD using fractional exhaled nitric oxide levels: a pilot study. International journal of chronic obstructive pulmonary disease. 2014;9:745.
30. Castaldi PJ, Benet M, Petersen H et al. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. Thorax. 2017 Nov 1;72(11):998-1006.
31. Radin G, Duan F, Billatos E, Snyder B, Stevenson C, Gatsonis C, O'Connor GT, Lenburg M, Washko G, Spira A. Characterizing Clinical And Imaging Phenotypes Of COPD Within The Decamp Consortium. InC22. COPD PHENOTYPES 2017 May (pp. A5002-A5002). American Thoracic Society.
32. Vazquez Guillamet R, Ursu O, Iwamoto G, Moseley PL, Oprea T. Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records. Health informatics journal. 2018 Dec;24(4):394-409.
33. Xavier RF, Pereira AC, Lopes AC, Cavalheri V, Pinto RM, Cukier A, Ramos EM, Carvalho CR. Identification of Phenotypes in People with COPD: Influence of Physical Activity, Sedentary Behaviour, Body Composition and Skeletal Muscle Strength. Lung. 2019 Feb 15;197(1):37-45.
34. de Torres JP, Marin JM, Martinez-Gonzalez C, de Lucas-Ramos P, Cosio B, Casanova C, COPD History Assessment In SpaiN (CHAIN) cohort. The importance of symptoms in the longitudinal variability of clusters in COPD patients: a validation study. Respirology. 2018 May;23(5):485-91.
35. Zarei S, Mirtar A, Morrow JD, Castaldi PJ, Belloni P, Hersh CP. Subtyping Chronic Obstructive Pulmonary Disease Using Peripheral Blood Proteomics. Chronic Obstructive Pulmonary Diseases. 2017;4(2):97.

36. Bafadhel M, Umar I, Gupta S, Raj JV, Vara DD, Entwisle JJ, Pavord ID, Brightling CE, Siddiqui S. The role of CT scanning in multidimensional phenotyping of COPD. Chest. 2011 Sep 1;140(3):634-42.
37. Lainez S, Court-Fortune I, Vercherin P, Falchero L, Didi T, Beynel P, Piperno D, Frappe E, Froudarakis M, Vergnon JM, Devouassoux G. Clinical ACO phenotypes: Description of a heterogeneous entity. Respiratory medicine case reports. 2019 Jan 1;28:100929.
38. Haghighi B, Choi S, Choi J, Hoffman EA, Comellas AP, Newell JD, Lee CH, Barr RG, Bleecker E, Cooper CB, Couper D. Imaging-based clusters in former smokers of the COPD cohort associate with clinical characteristics: the SubPopulations and intermediate outcome measures in COPD study (SPIROMICS). Respiratory research. 2019 Dec;20(1):153.
39. Karayama M, Inui N, Yasui H, Kono M, Hozumi H, Suzuki Y, Furuhashi K, Hashimoto D, Enomoto N, Fujisawa T, Nakamura Y. Clinical features of three-dimensional computed tomography-based radiologic phenotypes of chronic obstructive pulmonary disease. International journal of chronic obstructive pulmonary disease. 2019;14:1333.
40. Ning P, Guo YF, Sun TY, Zhang HS, Chai D, Li XM. Study of the clinical phenotype of symptomatic chronic airways disease by hierarchical cluster analysis and two-step cluster analyses. Zhonghua nei ke za zhi. 2016 Sep;55(9):679-83.
41. Rootmensen G, van Keimpema A, Zwinderman A, Sterk P. Clinical phenotypes of obstructive airway diseases in an outpatient population. Journal of Asthma. 2016 Nov 25;53(10):1026-32.
42. Fens N, van Rossum AG, Zanen P, van Ginneken B, van Klaveren RJ, Zwinderman AH, Sterk PJ. Subphenotypes of mild-to-moderate COPD by factor and cluster analysis of pulmonary function, CT imaging and breathomics in a population-based survey. COPD:

Journal of Chronic Obstructive Pulmonary Disease. 2013 Jun 1;10(3):277-85.

43. Bafadhel M, McKenna S, Terry S, Mistry V, Reid C, Haldar P, McCormick M, Haldar K, Kebadze T, Duvoix A, Lindblad K. Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers. American journal of respiratory and critical care medicine. 2011 Sep 15;184(6):662-71.

44. Cho MH, Washko GR, Hoffmann TJ, Criner GJ, Hoffman EA, Martinez FJ, Laird N, Reilly JJ, Silverman EK. Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation. Respiratory research. 2010 Dec 1;11(1):30.

45. Incalzi RA, Canonica GW, Scichilone N, Rizzoli S, Simoni L, Blasi F, STORICO study group. The COPD multi-dimensional phenotype: A new classification from the STORICO Italian observational study. PloS one. 2019;14(9).

46. Raherison C, Ouaalaya EH, Bernady A, Casteigt J, Nocent-Eijnani C, Falque L, Le Guillou F, Nguyen L, Ozier A, Molimard M. Comorbidities and COPD severity in a clinic-based cohort. BMC pulmonary medicine. 2018 Dec 1;18(1):117.

47. de Vries R, Dagelet YW, Spoor P, Snoey E, Jak PM, Brinkman P, Dijkers E, Bootsma SK, Elskamp F, De Jongh FH, Haarman EG. Clinical and inflammatory phenotyping by breathomics in chronic airway diseases irrespective of the diagnostic label. European Respiratory Journal. 2018 Jan 1;51(1).

48. Fingleton J, Huang K, Weatherall M, Guo Y, Ivanov S, Bruijnzeel P, Zhang H, Wang W, Beasley R, Wang C. Phenotypes of symptomatic airways disease in China and New Zealand. European Respiratory Journal. 2017 Dec 1;50(6):1700957.

49. Lee JH, Rhee CK, Kim K, Kim JA, Kim SH, Yoo KH, Kim WJ, Park YB, Park HY, Jung KS. Identification of subtypes in subjects with mild-to-moderate airflow limitation and its clinical and socioeconomic implications. International journal of chronic obstructive

pulmonary disease. 2017;12:1135.

50. Sekiya K, Nakatani E, Fukutomi Y, Kaneda H, Iikura M, Yoshida M, Takahashi K, Tomii K, Nishikawa M, Kaneko N, Sugino Y. Severe or life-threatening asthma exacerbation: patient heterogeneity identified by cluster analysis. Clinical & Experimental Allergy. 2016 Aug;46(8):1043-55.
51. Weatherall M, Travers J, Shirtcliffe PM, Marsh SE, Williams MV, Nowitz MR, Aldington S, Beasley R. Distinct clinical phenotypes of airways disease defined by cluster analysis. European Respiratory Journal. 2009 Oct 1;34(4):812-8.
52. Nikolaou V. Statistical analysis: a practical guide for psychiatrists. BJPsych Advances. 2016 Jul;22(4):251-9.
53. Mori Y, Kuroda M, Makino N. Nonlinear principal component analysis. In Nonlinear Principal Component Analysis and Its Applications 2016 (pp. 7-20). Springer, Singapore.
54. Lawley DN, Maxwell AE. Regression and factor analysis. Biometrika. 1973 Aug 1;60(2):331-8.
55. Joereskog KG. Statistical estimation in factor analysis. Almqvist & Wiksell; 1963.
56. Murtagh F, Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. Journal of classification. 2014 Oct 1;31(3):274-95.
57. Basagaña X, Barrera-Gómez J, Benet M, Antó JM, Garcia-Aymerich J. A framework for multiple imputation in cluster analysis. American journal of epidemiology. 2013 Apr 1;177(7):718-25.
58. Bourbeau J, Pinto LM, Benedetti A. Phenotyping of COPD: challenges and next steps. The Lancet Respiratory Medicine. 2014 Mar 1;2(3):172-4.
59. Biobank UK. https://www.ukbiobank.ac.uk/ (accessed Aug 15, 2019).

60. Pascoe S, Barnes N, Brusselle G, Compton C, Criner GJ, Dransfield MT, Halpin DM, Han MK, Hartley B, Lange P, Lettis S. Blood eosinophils and treatment response with triple and dual combination therapy in chronic obstructive pulmonary disease: analysis of the IMPACT trial. The Lancet Respiratory Medicine. 2019 Sep 1;7(9):745-56.
61. Sivapalan P, Lapperre TS, Janner J, Laub RR, Moberg M, Bech CS, Eklöf J, Holm FS, Armbruster K, Sivapalan P, Mosbech C. Eosinophil-guided corticosteroid therapy in patients admitted to hospital with COPD exacerbation (CORTICO-COP): a multicentre, randomised, controlled, open-label, non-inferiority trial. The Lancet Respiratory Medicine. 2019 Aug 1;7(8):699-709.
62. van Geffen WH, Slebos DJ, Herth FJ, Kemp SV, Weder W, Shah PL. Surgical and endoscopic interventions that reduce lung volume for emphysema: a systemic review and meta-analysis. The Lancet Respiratory Medicine. 2019 Apr 1;7(4):313-24.
63. Sun P, Ye R, Wang C, Bai S, Zhao L. Identification of proteomic signatures associated with COPD frequent exacerbators. Life sciences. 2019 Aug 1;230:1-9.
64. Pichl A, Sommer N, Bednorz M, Seimetz M, Hadzic S, Kuhnert S, Kraut S, Roxlau ET, Kojonazarov B, Wilhelm J, Gredic M. Riociguat for treatment of pulmonary hypertension in COPD: a translational study. European Respiratory Journal. 2019 Jun 1;53(6):1802445.
65. Pragman AA, Knutson KA, Gould TJ, Isaacson RE, Reilly CS, Wendt CH. Chronic obstructive pulmonary disease upper airway microbiota alpha diversity is associated with exacerbation phenotype: a case-control observational study. Respiratory research. 2019 Dec 1;20(1):114.
66. Kukol LV, Pupyshev SA. Determination of phenotypic characteristics of chronic obstructive lung disease in elderly patients. Advances in gerontology= Uspekhi gerontologii. 2019;32(3):445-50.

553 67. Pragman AA, Knutson KA, Gould TJ, Hodgson SW, Isaacson RE, Reilly CS, Wendt CH.
554 Chronic obstructive pulmonary disease upper airway microbiome is associated with
555 select clinical characteristics. PloS one. 2019;14(7).
556 68. Bak SH, Park HY, Nam JH, Lee HY, Lee JH, Sohn I, Chung MP. Predicting clinical outcome
557 with phenotypic clusters using quantitative CT fibrosis and emphysema features in
558 patients with idiopathic pulmonary fibrosis. PloS one. 2019;14(4).
559 69. Kneppers AE, Haast RA, Langen RC, Verdijk LB, Leermakers PA, Gosker HR, van Loon LJ,
560 Lainscak M, Schols AM. Distinct skeletal muscle molecular responses to pulmonary
561 rehabilitation in chronic obstructive pulmonary disease: a cluster analysis. Journal of
562 cachexia, sarcopenia and muscle. 2019 Apr;10(2):311-22.
563 70. Gedebjerg A, Szépligeti SK, Wackerhausen LM, Horváth-Puhó E, Dahl R, Hansen JG,
564 Sørensen HT, Nørgaard M, Lange P, Thomsen RW. Prediction of mortality in patients
565 with chronic obstructive pulmonary disease with the new Global Initiative for Chronic
566 Obstructive Lung Disease 2017 classification: a cohort study. The Lancet Respiratory
567 Medicine. 2018 Mar 1;6(3):204-12.
568 71. Merrill M, Roeder C, Butler M, Doran B, Stevens L, Goerg C, Kao D. Complex Heart
569 Failure Phenotypes Differ in Response to Medical Therapy and Exercise Training.
570 Circulation. 2018 Nov 6;138(Suppl_1):A16910.
571 72. El Boueiz AR, Chang Y, Cho MH, DeMeo DL, Dy J, Silverman EK, Castaldi P. Machine
572 Learning Prediction of 5-Year Progression of FEV1 in the COPDGene Study. InD101.
573 MECHANISTIC AND TRANSLATIONAL STUDIES IN COPD 2018 May (pp. A7430-A7430).
574 American Thoracic Society.
575 73. Fang L, Gao P, Bao H, Tang X, Wang B, Feng Y, Cong S, Juan J, Fan J, Lu K, Wang N.
576 Chronic obstructive pulmonary disease in China: a nationwide prevalence study. The

Lancet Respiratory Medicine. 2018 Jun 1;6(6):421-30.

74. Koo HK, Vasilescu DM, Booth S, Hsieh A, Katsamenis OL, Fishbane N, Elliott WM, Kirby M, Lackie P, Sinclair I, Warner JA. Small airways disease in mild and moderate chronic obstructive pulmonary disease: a cross-sectional study. The Lancet Respiratory Medicine. 2018 Aug 1;6(8):591-602.

75. Liang X, Sha Q, Rho Y, Zhang S. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. Genetic epidemiology. 2018 Jun;42(4):344-53.

76. Kilk K, Aug A, Ottas A, Soomets U, Altraja S, Altraja A. Phenotyping of chronic obstructive pulmonary disease based on the integration of metabolomes and clinical characteristics. International journal of molecular sciences. 2018 Mar;19(3):666.

77. Le Rouzic O, Roche N, Cortot AB, Tillie-Leblond I, Masure F, Perez T, Boucot I, Hamouti L, Ostinelli J, Pribil C, Poutchnine C. Defining the “frequent exacerbator” phenotype in COPD: a hypothesis-free approach. Chest. 2018 May 1;153(5):1106-15.

78. Hall M, Dondo TB, Yan AT, Mamas MA, Timmis AD, Deanfield JE, Jernberg T, Hemingway H, Fox KA, Gale CP. Multimorbidity and survival for patients with acute myocardial infarction in England and Wales: Latent class analysis of a nationwide population-based cohort. PLoS medicine. 2018 Mar;15(3).

79. Das N, Topalovic M, Janssens W. Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential. Current opinion in pulmonary medicine. 2018 Mar 1;24(2):117-23.

80. Merchant R, Szefler SJ, Bender BG, Tuffli M, Barrett MA, Gondalia R, Kaye L, Van Sickle D, Stempel DA. Impact of a digital health intervention on asthma resource utilization. World Allergy Organization Journal. 2018 Dec 1;11(1):28.

81. Christenson S, Bolourchi S, Huffnagle G, Erb-Downward J, Hanauer G, Saetta M, Rabe K,

Martinez FJ, Woodruff PG. Molecular phenotyping of chronic bronchitis: mucin and inflammatory gene expression heterogeneity in COPD.

82. Kästle M, Bartel S, Geillinger-Kästle K, Irmler M, Beckers J, Ryffel B, Eickelberg O, Krauss-Etschmann S. micro RNA cluster 106a~ 363 is involved in T helper 17 cell differentiation. Immunology. 2017 Nov;152(3):402-13.

83. Fouda MA, Alhamad EH, Al-Hajjaj MS, Shaik SA, Alboukai AA, Al-Kassimi FA. A study of chronic obstructive pulmonary disease-specific causes of osteoporosis with emphasis on the emphysema phenotype. Annals of thoracic medicine. 2017 Apr;12(2):101.

84. Chalmers JD. Bronchiectasis: phenotyping a complex disease. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2017 Mar 15;14(sup1):S12-8.

85. Hirai K, Shirai T, Suzuki M, Akamatsu T, Suzuki T, Hayashi I, Yamamoto A, Akita T, Morita S, Asada K, Tsuji D. A clustering approach to identify and characterize the asthma and chronic obstructive pulmonary disease overlap phenotype. Clinical & Experimental Allergy. 2017 Nov;47(11):1374-82.

86. Haldar K, Bafadhel M, Lau K, Berg A, Kwambana B, Kebadze T, Ramsheh MY, Barker B, Haldar P, Johnston S, Ketley JM. Microbiome balance in sputum determined by PCR stratifies COPD exacerbations and shows potential for selective use of antibiotics. PLoS One. 2017;12(8).

87. Çolak Y, Afzal S, Nordestgaard BG, Vestbo J, Lange P. Prognosis of asymptomatic and symptomatic, undiagnosed COPD in the general population in Denmark: a prospective cohort study. The Lancet Respiratory Medicine. 2017 May 1;5(5):426-34.

88. Maddocks M, Nolan CM, Man WD, Polkey MI, Hart N, Gao W, Rafferty GF, Moxham J, Higginson IJ. Neuromuscular electrical stimulation to improve exercise capacity in patients with severe COPD: a randomised double-blind, placebo-controlled trial. The

625 Lancet Respiratory Medicine. 2016 Jan 1;4(1):27-36.
626 89. Lange P, Çolak Y, Ingebrigtsen TS, Vestbo J, Marott JL. Long-term prognosis of asthma,
627 chronic obstructive pulmonary disease, and asthma-chronic obstructive pulmonary
628 disease overlap in the Copenhagen City Heart study: a prospective population-based
629 analysis. The Lancet Respiratory Medicine. 2016 Jun 1;4(6):454-62.
630 90. Sato S, Tanino Y, Misa K, Fukuhara N, Nikaido T, Uematsu M, Fukuhara A, Wang X, Ishida
631 T, Munakata M. Identification of clinical phenotypes in idiopathic interstitial pneumonia
632 with pulmonary emphysema. Internal Medicine. 2016 Jun 15;55(12):1529-35.
633 91. Morélot-Panzini C, Gilet H, Aguilaniu B, Devillier P, Didier A, Perez T, Pignier C, Arnould
634 B, Similowski T. Real-life assessment of the multidimensional nature of dyspnoea in
635 COPD outpatients. European Respiratory Journal. 2016 Jun 1;47(6):1668-79.
636 92. Roche O, Deguiz ML, Tiana M, Galiana-Ribote C, Martinez-Alcazar D, Rey-Serra C, Ranz-
637 Ribeiro B, Casitas R, Galera R, Fernández-Navarro I, Sánchez-Cuéllar S. Identification of
638 non-coding genetic variants in samples from hypoxemic respiratory disease patients that
639 affect the transcriptional response to hypoxia. Nucleic acids research. 2016 Nov
640 2;44(19):9315-30.
641 93. Martínez-García MÁ, Vendrell M, Girón R, Máiz-Carro L, de la Rosa Carrillo D, de Gracia J,
642 Olveira C. The Multiple Faces of Non–Cystic Fibrosis Bronchiectasis. A Cluster Analysis
643 Approach. Annals of the American Thoracic Society. 2016 Sep;13(9):1468-75.
644 94. Batista-Navarro R, Carter J, Ananiadou S. Argo: enabling the development of bespoke
645 workflows and services for disease annotation. Database. 2016 Jan 1;2016.
646 95. Labuzzetta CJ, Antonio ML, Watson PM, Wilson RC, Laboissonniere LA, Trimarchi JM,
647 Genc B, Ozdinler PH, Watson DK, Anderson PE. Complementary feature selection from
648 alternative splicing events and gene expression for phenotype prediction.

649 Bioinformatics. 2016 Sep 1;32(17):i421-9.
650 96. Kaluarachchi MR, Boulangé CL, Garcia-Perez I, Lindon JC, Minet EF. Multiplatform serum
651 metabolic phenotyping combined with pathway mapping to identify biochemical
652 differences in smokers. Bioanalysis. 2016 Oct;8(19):2023-43.
653 97. Obeidat ME, Hao K, Bossé Y, Nickle DC, Nie Y, Postma DS, Laviolette M, Sandford AJ,
654 Daley DD, Hogg JC, Elliott WM. Molecular mechanisms underlying variations in lung
655 function: a systems genetics analysis. The Lancet Respiratory Medicine. 2015 Oct
656 1;3(10):782-95.
657 98. Kim S, Herazo-Maya JD, Kang DD, Juan-Guardela BM, Tedrow J, Martinez FJ, Sciurba FC,
658 Tseng GC, Kaminski N. Integrative phenotyping framework (iPF): integrative clustering of
659 multiple omics data identifies novel lung disease subphenotypes. BMC genomics. 2015
660 Dec 1;16(1):924.
661 99. Huebenthal M, Hemmrich-Stanisak G, Degenhardt F, Szymczak S, Du Z, Elsharawy A,
662 Keller A, Schreiber S, Franke A. Sparse modeling reveals miRNA signatures for
663 diagnostics of inflammatory bowel disease. PloS one. 2015;10(10).
664 100. Lee JH, Cho MH, McDonald ML, Hersh CP, Castaldi PJ, Crapo JD, Wan ES, Dy JG,
665 Chang Y, Regan EA, Hardin M. Phenotypic and genetic heterogeneity among subjects
666 with mild airflow obstruction in COPDGene. Respiratory medicine. 2014 Oct
667 1;108(10):1469-80.
668 101. Uzun S, Djamin RS, Kluytmans JA, Mulder PG, van't Veer NE, Ermens AA, Pelle AJ,
669 Hoogsteden HC, Aerts JG, van der Eerden MM. Azithromycin maintenance treatment in
670 patients with frequent exacerbations of chronic obstructive pulmonary disease
671 (COLUMBUS): a randomised, double-blind, placebo-controlled trial. The Lancet
672 Respiratory Medicine. 2014 May 1;2(5):361-8.

673 102. Brightling CE, Bleecker ER, Panettieri Jr RA, Bafadhel M, She D, Ward CK, Xu X, Birrell
674 C, van der Merwe R. Benralizumab for chronic obstructive pulmonary disease and
675 sputum eosinophilia: a randomised, double-blind, placebo-controlled, phase 2a study.
676 The Lancet Respiratory Medicine. 2014 Nov 1;2(11):891-901.
677 103. Kon SS, Canavan JL, Jones SE, Nolan CM, Clark AL, Dickson MJ, Haselden BM, Polkey
678 MI, Man WD. Minimum clinically important difference for the COPD Assessment Test: a
679 prospective analysis. The Lancet Respiratory Medicine. 2014 Mar 1;2(3):195-203.
680 104. Köhnlein T, Windisch W, Köhler D, Drabik A, Geiseler J, Hartl S, Karg O, Laier-
681 Groeneveld G, Nava S, Schönhofer B, Schucher B. Non-invasive positive pressure
682 ventilation for the treatment of severe stable chronic obstructive pulmonary disease: a
683 prospective, multicentre, randomised, controlled clinical trial. The Lancet Respiratory
684 Medicine. 2014 Sep 1;2(9):698-705.
685 105. Jones RC, Price D, Ryan D, Sims EJ, von Ziegenweidt J, Mascarenhas L, Burden A,
686 Halpin DM, Winter R, Hill S, Kearney M. Opportunities to diagnose chronic obstructive
687 pulmonary disease in routine care in the UK: a retrospective study of a clinical cohort.
688 The Lancet Respiratory Medicine. 2014 Apr 1;2(4):267-76.
689 106. Zheng JP, Wen FQ, Bai CX, Wan HY, Kang J, Chen P, Yao WZ, Ma LJ, Li X, Raiteri L,
690 Sardina M. Twice daily N-acetylcysteine 600 mg for exacerbations of chronic obstructive
691 pulmonary disease (PANTHEON): a randomised, double-blind placebo-controlled trial.
692 The Lancet Respiratory Medicine. 2014 Mar 1;2(3):187-94.
693 107. Corhay JL, Schleich F, Louis R. Phenotypes in chronic obstructive pulmonary disease.
694 Revue medicale de Liege. 2014;69(7-8):415-21.
695 108. Moore WC, Hastie AT, Li X, Li H, Busse WW, Jarjour NN, Wenzel SE, Peters SP,
696 Meyers DA, Bleecker ER, Heart N. Sputum neutrophil counts are associated with more

severe asthma phenotypes using cluster analysis. Journal of Allergy and clinical immunology. 2014 Jun 1;133(6):1557-63.

109. Qiao D, Cho MH, Fier H, Bakke PS, Gulsvik A, Silverman EK, Lange C. On the simultaneous association analysis of large genomic regions: a massive multi-locus association test. Bioinformatics. 2014 Jan 15;30(2):157-64.

110. DiSantostefano RL, Li H, Hinds D, Galkin DV, Rubin DB. Risk of pneumonia with inhaled corticosteroid/long-acting β2 agonist therapy in chronic obstructive pulmonary disease: a cluster analysis. International journal of chronic obstructive pulmonary disease. 2014;9:457.

111. Vogelmeier CF, Bateman ED, Pallante J, Alagappan VK, D'Andrea P, Chen H, Banerji D. Efficacy and safety of once-daily QVA149 compared with twice-daily salmeterol–fluticasone in patients with chronic obstructive pulmonary disease (ILLUMINATE): a randomised, double-blind, parallel group study. The Lancet Respiratory Medicine. 2013 Mar 1;1(1):51-60.

112. Franciosi LG, Diamant Z, Banner KH, Zuiker R, Morelli N, Kamerling IM, de Kam ML, Burggraaf J, Cohen AF, Cazzola M, Calzetta L. Efficacy and safety of RPL554, a dual PDE3 and PDE4 inhibitor, in healthy volunteers and in patients with asthma or chronic obstructive pulmonary disease: findings from four clinical trials. The lancet Respiratory medicine. 2013 Nov 1;1(9):714-27.

113. Decramer ML, Chapman KR, Dahl R, Frith P, Devouassoux G, Fritscher C, Cameron R, Shoaib M, Lawrence D, Young D, McBryan D. Once-daily indacaterol versus tiotropium for patients with severe chronic obstructive pulmonary disease (INVIGORATE): a randomised, blinded, parallel-group study. The Lancet Respiratory medicine. 2013 Sep 1;1(7):524-33.

114. Dransfield MT, Bourbeau J, Jones PW, Hanania NA, Mahler DA, Vestbo J, Wachtel A, Martinez FJ, Barnhart F, Sanford L, Lettis S. Once-daily inhaled fluticasone furoate and vilanterol versus vilanterol only for prevention of exacerbations of COPD: two replicate double-blind, parallel-group, randomised controlled trials. The Lancet Respiratory Medicine. 2013 May 1;1(3):210-23.

115. Wedzicha JA, Decramer M, Ficker JH, Niewoehner DE, Sandström T, Taylor AF, D'Andrea P, Arrasate C, Chen H, Banerji D. Analysis of chronic obstructive pulmonary disease exacerbations with the dual bronchodilator QVA149 compared with glycopyrronium and tiotropium (SPARK): a randomised, double-blind, parallel-group study. The lancet Respiratory medicine. 2013 May 1;1(3):199-209.

116. Rabe KF, Fabbri LM, Israel E, Kögler H, Riemann K, Schmidt H, Glaab T, Vogelmeier CF. Effect of ADRB2 polymorphisms on the efficacy of salmeterol and tiotropium in preventing COPD exacerbations: a prespecified substudy of the POET-COPD trial. The Lancet Respiratory Medicine. 2014 Jan 1;2(1):44-53.

117. Siedlinski M, Tingley D, Lipman PJ, Cho MH, Litonjua AA, Sparrow D, Bakke P, Gulsvik A, Lomas DA, Anderson W, Kong X. Dissecting direct and indirect genetic effects on chronic obstructive pulmonary disease (COPD) susceptibility. Human genetics. 2013 Apr 1;132(4):431-41.

118. Gouzi F, Abdellaoui A, Molinari N, Pinot E, Ayoub B, Laoudj-Chenivesse D, Cristol JP, Mercier J, Hayot M, Préfaut C. Fiber atrophy, oxidative stress, and oxidative fiber reduction are the attributes of different phenotypes in chronic obstructive pulmonary disease patients. Journal of Applied Physiology. 2013 Dec 15;115(12):1796-805.

119. Shaykhiev R, Sackrowitz R, Fukui T, Zuo WL, Chao IW, Strulovici-Barel Y, Downey RJ, Crystal RG. Smoking-induced CXCL14 expression in the human airway epithelium links

chronic obstructive pulmonary disease to lung cancer. American journal of respiratory cell and molecular biology. 2013 Sep;49(3):418-25.

120. Carolan BJ, Sutherland ER. Clinical phenotypes of chronic obstructive pulmonary disease and asthma: recent advances. Journal of allergy and clinical immunology. 2013 Mar 1;131(3):627-34.

121. Toraldo DM, Minelli M, De Nuccio F, Nicolardi G. Chronic obstructive pulmonary disease phenotype desaturator with hypoxic vascular remodelling and pulmonary hypertension obtained by cluster analysis. Multidisciplinary respiratory medicine. 2012 Dec;7(1):39.

122. Travers J, Weatherall M, Fingleton J, Beasley R. Towards individualised medicine for airways disease: identifying clinical phenotype groups. European Respiratory Journal. 2012 Apr 1;39(4):1033-4.

123. Toraldo DM, De Nuccio F, Gaballo A, Nicolardi G. Use of cluster analysis to describe desaturator phenotypes in COPD: correlations between pulmonary function tests and nocturnal oxygen desaturation. International journal of chronic obstructive pulmonary disease. 2011;6:551.

124. Fingleton J, Weatherall M, Beasley R. Towards individualised treatment in COPD.

125. Shirtcliffe P, Weatherall M, Travers J, Beasley R. The multiple dimensions of airways disease: targeting treatment to clinical phenotypes. Current opinion in pulmonary medicine. 2011 Mar 1;17(2):72-8.

126. Sharma S, Miller DP, Emmett A, Li H. Cluster Analysis For The Identification And Replication Of Distinct Subject Clusters From COPD Clinical Trials. InA41. CHRONIC OBSTRUCTIVE PULMONARY DISEASE EXACERBATIONS: EPIDEMIOLOGY AND OUTCOMES 2010 May (pp. A1501-A1501). American Thoracic Society.

127. Jo KW, Ra SW, Chae EJ, Seo JB, Kim NK, Lee JH, Kim EK, Lee YK, Kim TH, Huh JW, Kim WJ. Three phenotypes of obstructive lung disease in the elderly. The International journal of tuberculosis and lung disease. 2010 Nov 1;14(11):1481-8.

128. Sobradillo P, Garcia-Aymerich J, Agusti A. Clinical phenotypes of COPD. Archivos de bronconeumologia. 2010 Dec;46:8-11.

129. Weatherall M, Shirtcliffe P, Travers J, Beasley R. Use of cluster analysis to define COPD phenotypes. European respiratory journal. 2010 Sep 1;36(3):472-4.

130. Paoletti M, Camiciottoli G, Meoni E, Bigazzi F, Cestelli L, Pistolesi M, Marchesi C. Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes. Journal of biomedical informatics. 2009 Dec 1;42(6):1013-21.

131. Pistolesi M, Camiciottoli G, Paoletti M, Marmai C, Lavorini F, Meoni E, Marchesi C, Giuntini C. Identification of a predominant COPD phenotype in clinical practice. Respiratory medicine. 2008 Mar 1;102(3):367-76.

132. Patel BD, Coxson HO, Pillai SG, Agusti AG, Calverley PM, Donner CF, Make BJ, Muller NL, Rennard SI, Vestbo J, Wouters EF. Airway wall thickening and emphysema show independent familial aggregation in chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine. 2008 Sep 1;178(5):500-5.

133. Kodavanti UP, Schladweiler MC, Ledbetter AD, Ortuno RV, Suffia M, Evansky P, Richards JH, Jaskot RH, Thomas R, Karoly E, Huang YC. The Spontaneously Hypertensive Rat: An Experimental Model of Sulfur Dioxide–Induced Airways Disease. Toxicological Sciences. 2006 Nov 1;94(1):193-205.

134. Wardlaw AJ, Silverman M, Siva R, Pavord ID, Green R. Multi-dimensional phenotyping: towards a new taxonomy for airway disease. Clinical & Experimental

Allergy. 2005 Oct;35(10):1254-62.

135. Hackett NR, Heguy A, Harvey BG, O'Connor TP, Luettich K, Flieder DB, Kaplan R, Crystal RG. Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. American journal of respiratory cell and molecular biology. 2003 Sep;29(3):331-43.

Table 1. 2019 GOLD classification of COPD phenotypes

	Symptoms	
Moderate/severe exacerbation history	mMRC 0-1 and CAT<10	mMRC≥2 and CAT≥10
≥2 or ≥1 leading to hospital admission	C	D
0 or 1 not leading to hospital admission	A	B

mMRC, modified Medical Research Council dyspnea questionnaire; CAT, COPD assessment test

Table 2. Summary of studies using clustering analysis to identify COPD phenotypes used in the systematic analysis

First author and year of publication	Sample size (i.e., number of patients) contributing to cluster analysis	Name of cohort and study design	Population characteristics and setting(s)	COPD phenotypes identified	Clinical outcome(s) used for validation
Yoon et al. (2019) [9]	1,195	Korea COPD subgroup study (KOCOSS), retrospective observational multi-centre longitudinal cohort	Patients with COPD evaluated at 6-month intervals by experienced pulmonologists at university hospitals	1. Putative asthma-COPD overlap 2. Mild COPD 3. Moderate COPD 4. Severe COPD	Acute exacerbation
Pikoula et al. (2019) [6]	30,961	CALIBER1, observational prospective longitudinal cohort	Patients who a) were 35 years or older, b) had been registered for at least one year in primary care practice, c) had at least one diagnostic code of COPD	1. Anxiety/depression 2. Severe airflow obstruction and frailty 3. Cardiovascular disease and diabetes 4. Obesity/atopy 5. Low prevalence of comorbidities	Rate of severe or moderate acute COPD exacerbations, respiratory and cardiovascular related mortality
Kim et al. (2018) [10]	1,676	The Asian Network for Obstructive Lung Disease (ANOLD) international multi-centre	Patients of Asian ethnicity, over 40 years old with FEV1/FVC < 0.7 assessed at pulmonary clinics	1. Worse lung function but fewer symptoms 2. Worse lung function with more symptoms and most frequent exacerbations,	Exacerbations and quality of life

First author and year of publication	Sample size (i.e., number of patients) contributing to cluster analysis	Name of cohort and study design	Population characteristics and setting(s)	COPD phenotypes identified	Clinical outcome(s) used for validation
		observational cross-sectional prospective cohort		faster FEV1 decline and greatest SGRQ decline 3. Mild severity but higher BMI	
Kim et al. (2017) [11]	272	COPD in dusty areas (CODA) observational longitudinal prospective cohort	Patients over 40 years old with FEV1/FVC < 0.7 living near cement plants who were evaluated at enrolment and at a 1-year follow-up at university hospitals	1. Younger patients with fewer symptoms and exacerbations and mild airflow obstruction 2. Patients with additional symptoms and moderate airflow obstruction and more exacerbations requiring hospitalization 3. More female patients, additional symptoms and mild airflow obstruction and modest frequency of exacerbations requiring hospitalization	Exacerbations and quality of life
Burgel et. al (2017) [8]	2,409	Three French/Belgian COPD cohorts: a) the initiatives BPCO observational	Patients with stable COPD assessed at university hospitals	1. Older patients with high rates of cardiovascular comorbidities and diabetes, but less severe respiratory	3-year all-cause mortality

First author and year of publication	Sample size (i.e., number of patients) contributing to cluster analysis	Name of cohort and study design	Population characteristics and setting(s)	COPD phenotypes identified	Clinical outcome(s) used for validation
		multi-centre prospective cross-sectional cohort, b) the CPHG2 prospective observational cohort, c) the Leuven observational cross-sectional cohort. An independent cohort (the 3CIA3 initiative) was also used for validation.		disease 2. Moderate to severe respiratory disease and low rate of comorbidities 3. Older patients with high prevalence of comorbidities and obesity 4. Very severe respiratory disease with low rates of cardiovascular comorbidities and diabetes 5. Mild respiratory disease and low rates of comorbidities	
Peters et al. (2017) [14]	619	Two interventional cohorts: a) 1-year follow-up treatment as usual (TAU), b) 12-week pulmonary rehabilitation (PR) program	Two groups of patients: 160 out-patients with COPD treated as usual (TAU) and 459 patients with pulmonary rehabilitation (PR) at a university medical centre	1. Moderate COPD, low impact on health status (adaptive phenotype) 2. Severe COPD, moderate impact on health status (adaptive) 3. Moderate COPD, high impact on health status (non-adaptive)	Response to treatment (TAU vs PR)

First author and year of publication	Sample size (i.e., number of patients) contributing to cluster analysis	Name of cohort and study design	Population characteristics and setting(s)	COPD phenotypes identified	Clinical outcome(s) used for validation
Chubachi et al. (2016) [17]	311	The Keio COPD Comorbidity Research (K-CCR) observational, prospective cohort	COPD patients with complete comorbidities data with a 2-year follow-up assessed at Keio University and its affiliated hospitals	1. Less comorbidity 2. Malignancy 3. Metabolic and cardiovascular 4. Gastroesophageal reflux disease (GERD) and psychological 5. Underweight and anaemic	Health-related quality of life (e.g. SGRQ, CAT, SF-36)
Fingleton et al. (2015) [19]	389	A 3-phase cross-sectional study; phase 1 (sample selection), phase 2 (phenotyping), phase 3 (interventional study to assess treatment responsiveness)	Patients with symptoms of wheezing and breathlessness in the last 12 months who completed phase 2 with no missing data	1. Moderate to severe atopic asthma 2. Asthma-COPD overlap 3. Obese/comorbid 4. Mild atopic asthma 5. Mild intermittent	Response to treatment (inhaled β-agonist, antimuscarinic, corticosteroid)
Chen et al. (2014) [16]	332	Observational prospective longitudinal cohort	Men with COPD diagnosed at university hospital	1. Young patients with mild airflow obstruction, few symptoms and infrequent severe exacerbations 2. Older patients with mild airflow obstruction, few symptoms, infrequent severe exacerbations but higher mortality 3. Older patients with	Mortality

First author and year of publication	Sample size (i.e., number of patients) contributing to cluster analysis	Name of cohort and study design	Population characteristics and setting(s)	COPD phenotypes identified	Clinical outcome(s) used for validation
				moderate respiratory disease, dyspnoea, history of severe exacerbations and underweight 4. Patients with severe airflow obstruction, many symptoms and infrequent severe exacerbations 5. Patients with severe airflow obstruction, many symptoms and frequent severe exacerbations and high mortality	
Castaldi et al. (2014) [7]	8,288	The Genetic Epidemiology of COPD (COPDGene) study observational cross-sectional prospective cohort	Former and current smokers with or without COPD	1. No/mild obstruction and minimal emphysema 2. Mild upper zone emphysema predominant 3. Airway disease predominant 4. Severe emphysema	Exacerbations, dyspnoea, COPD-associated genetic variants
Altenburg et al. (2012) [18]	65	An interventional prospective cohort	Patients with COPD participating in a pulmonary rehabilitation (PR) program at a university medical centre	1. Worse lung function, quadriceps force but better response to exercise training 2. Better lung function and exercise capacity but less	Improvement in exercise capacity

First author and year of publication	Sample size (i.e., number of patients) contributing to cluster analysis	Name of cohort and study design	Population characteristics and setting(s)	COPD phenotypes identified	Clinical outcome(s) used for validation
				response to exercise training	
Burgel et al. (2010) [12]	322	The initiatives BPCO observational multi-centre prospective cross-sectional cohort	Patients with stable COPD assessed at 17 pulmonary units in university hospitals	1. Young patients with severe respiratory disease 2. Older patients with mild airflow limitation and mild comorbidities 3. Young patients with moderate to severe airflow limitation, few comorbidities 4. Older patients with moderate to severe airflow limitation and high prevalence of cardiovascular comorbidities	All-cause mortality
Burgel et al. (2012) [13]	527	Two cohorts: the Leuven observational cross-sectional cohort (374 patients) and the NELSON randomized lung cancer screening study (153 patients)	Stable COPD patients assessed at university hospitals' COPD outpatient clinics	1. Young patients with severe respiratory disease and low prevalence of cardiovascular comorbidities 2. Older patients with less severe airflow limitation, obese, high prevalence of diabetes and cardiovascular comorbidities 3. Mild to moderate airflow limitation,	All-cause mortality

First author and year of publication	Sample size (i.e., number of patients) contributing to cluster analysis	Name of cohort and study design	Population characteristics and setting(s)	COPD phenotypes identified	Clinical outcome(s) used for validation
				absent or mild emphysema and dyspnoea, normal nutritional status, limited comorbidities	
Garcia-Aymerich et al. (2011) [15]	342	An observational, prospective cross-sectional cohort	COPD patients hospitalized due to COPD exacerbation in teaching hospitals	1. Severe respiratory COPD 2. Moderate respiratory COPD 3. Systemic COPD (high prevalence of cardiovascular comorbidities)	Hospitalizations and all-cause mortality

[1]CALIBER: A database of electronic health records from three national sources; The Clinical Practice Research Datalink (CPRD), Hospital Episode Statistics (HES), and cause-specific mortality from the Office for National Statistics (ONS)
[2]CPHG: The French College of General Hospital Respiratory Physicians
[3]3CIA: COPD Cohorts Collaborative International Assessment
[4]NZRHS: New Zealand Respiratory Health Survey

Table 3. Summary of studies excluded from the systematic analysis

First author and year of publication	Type and purpose of study	Main findings	COPD phenotypes	Reason for exclusion
Pascoe et al. (2019) [60]	A randomized parallel group clinical trial aimed to model the relationships between eosinophil counts, smoking and treatment response to inhaled corticosteroids (ICS), and their interactions, including outcomes other than exacerbations.	Results showed that assessment of blood eosinophil count and smoking status has the potential to optimize ICS use in clinical practice in patients with COPD and a history of exacerbations.	Not applicable	Not relevant to machine learning methods under study
Sivapalan et al. (2019) [61]	A randomized controlled non-inferiority trial aimed to determine whether an algorithm based on blood eosinophil counts could safely reduce systemic corticosteroid exposure in patients admitted to hospital with acute exacerbations of COPD	Results showed that eosinophil-guided therapy was non-inferior compared with standard care for the number of days alive and out of hospital, and reduced the duration of systemic corticosteroid exposure,	Not applicable	Not relevant to COPD phenotyping
van Geffen et al. (2019) [62]	A systematic review and meta-analysis aimed to evaluate the effects of volume reduction in the treatment of severe emphysema	Results showed that lung volume reduction in patients with severe emphysema on maximal medical treatment has clinically meaningful benefits	Not applicable	Not relevant to COPD phenotyping
Sun et al. (2019) [63]	A cross-sectional study designed to detect proteins that were differentially abundant in COPD frequent exacerbators and assess whether those expression profiles are unique among COPD patients	Bioinformatics analyses of proteome indicated that the immune network for IgA production and the phenylalanine metabolism pathway were associated with frequent exacerbations	Not applicable	Not relevant with the machine learning methods under study
Pichl et al. (2019) [64]	A retrospective observational study investigated the treatment effect of riociguat and analysed the effect of	Data showed that riociguat may be beneficial for treatment of PH-COPD	Not applicable	Not relevant with the machine learning methods under study

	riociguat treatment on pulmonary hypertension (PH) in single patients with PH-COPD			
Pragman et al. (2019) [65]	A case-control observational study aimed to determine key features that differentiate the oral and sputum microbiota of frequent exacerbators (FEs) from the microbiota of infrequent exacerbators (IEs) during periods of clinical stability	Data showed that the frequent exacerbator phenotype is associated with decreased alpha diversity, beta-diversity clustering, and changes in taxonomic abundance	Not applicable	Not relevant with machine learning methods under study
Xavier et al. (2019) [33]	An observational cross-sectional study aiming to investigate COPD phenotypes according to their levels of physical activity and sedentary behaviour, as well as body composition and skeletal muscle strength	Cluster analysis identified three distinct COPD phenotypes	1) more physically active, less sedentary and had better body composition and lower ADO index, 2) older, less physically active, more sedentary having a higher dyspnoea and obstruction (ADO) index, 3) worse HRQoL, clinical control and body composition, less physically active, more sedentary having a higher ADO index	COPD phenotypes were not validated with clinical meaningful outcomes
Incalzi et al. (2019) [45]	The STORICO Italian observational study aiming to describe multi-dimensional COPD phenotypes	Machine learning methods used to identify five COPD phenotypes	1) Mild COPD: no night-time symptoms and the best health status in terms of quality of life, quality of sleep, level of depression and anxiety, 2) Mild emphysematous: prevalent dyspnea in the early-morning and daytime, 3) Severe bronchitic: nocturnal and diurnal cough and	COPD phenotypes were not validated with clinical meaningful outcomes

			phlegm, 4) Severe emphysematous: nocturnal and diurnal dyspnea, 5) Severe mixed COPD: higher frequency of symptoms during 24h and worst quality of life, of sleep and highest levels of depression and anxiety.	
Lainez et al. (2019) [37]	A retrospective study aiming to identify asthma and COPD overlap (ACO) phenotypes	Cluster analysis identified four ACO phenotypes	1) overweighed heavy smokers, with an early onset and a severe disease, 2) similar patients, with a late onset, 3) and 4) slighter smokers, presenting a moderate disease, with early and late onset respectively	ACO phenotypes were not validated with clinical meaningful outcomes
Kukol et al. (2019) [66]	A cross-sectional study aiming to identify COPD phenotypes of elderly patients	Cluster analysis identified different COPD phenotypes for men and women	Not applicable	COPD phenotypes were not validated with clinical meaningful outcomes
Pragman et al. (2019) [67]	A genetic study aiming to determine features that differentiate the oral, nasal, and sputum microbiome among subjects with stable COPD	Data showed associations between anatomic site and bacterial biomass, Shannon diversity, and β-diversity.	Not applicable	Not relevant to COPD phenotyping
Haghighi et al. (2019) [38]	A multi-center cross-sectional study aiming to identify COPD phenotypes using Quantitative computed tomographic (QCT) imaging	Imaging-based cluster analysis identified four possible COPD phenotypes	1) asymptomatic and showed relatively normal airway structure and lung function except airway wall thickening and moderate emphysema, 2) obese females showed an increase of tissue fraction at inspiration,	COPD phenotypes were not validated with clinical meaningful outcomes

			minimal emphysema, and the lowest progression rate of emphysema, 3) older males showed small airway narrowing and a decreased tissue fraction at expiration, both indicating air-trapping, 4) lean males were likely to be severe COPD subjects showing the highest progression rate of emphysema	
Bak et al. (2019) [68]	A retrospective observational study aimed to assess prognostic impact among identified clusters in patient with idiopathic pulmonary fibrosis (IPF) and evaluate the impact of fibrosis and emphysema on lung function	Cluster analysis identified distinct phenotypes, which predicted prognosis of clinical outcome	Not applicable	Not relevant to COPD phenotyping
Karayama et al. (2019) [39]	A cross-sectional study aimed to identify novel COPD phenotypes using radiologic data	Cluster analysis identified four COPD phenotypes	1) mild emphysema with severe airway changes, severe airflow limitation, and high exacerbation risk, 2) mild emphysema with moderate airway changes, mild airflow limitation, and mild dyspnea, 3) severe emphysema with moderate airway changes, severe airflow limitation, and increased dyspnea, 4) moderate emphysema with mild airway changes, mild airflow limitation, low exacerbation risk, and	COPD phenotypes were not validated with clinical meaningful outcomes

			mild dyspnea	
Kneppers et al (2019) [69]	A prospective observational study aimed to assess skeletal muscle molecular responses to Pulmonary rehabilitation (PR) in COPD patients	Cluster analysis identified patient groups with distinct skeletal muscle molecular responses to rehabilitation	Not applicable	COPD phenotypes were not validated with clinical meaningful outcomes
de Torres et al. (2018) [34]	A prosepctive observational study aimed to evaluate the 2-year cluster variability in stable COPD patients.	Data showed that after 2 years of follow-up, most of the COPD patients maintained their cluster assignment	1) younger age, mild airway limitation, few symptoms, 2) intermediate (clinical characteristics between clusters 1 and 3), 3) older age, severe airway limitation and highly symptomatic	COPD phenotypes were not validated with clinical meaningful outcomes
Gedebjerg et al. (2018) [70]	A prospective observational study aimed to establish the predictive ability of the GOLD 2017 classification, compared with earlier classifications, for all-cause and respiratory mortality	Data showed that the new GOLD 2017 ABCD classification does not predict all-cause and respiratory mortality more accurately than the previous GOLD systems from 2007 and 2011	Not applicable	Not relevant to COPD phenotyping and to machine learning methods under study
Merrill et al. (2018) [71]	Data from two randomized clinical trials aimed to investigate the response to specific interventions according to heart failure (HF) phenotype	Response to treatments such as exercise training and spironolactone varies among complex HF phenotypes	Not applicable	Not relevant to COPD phenotyping
El Boueiz (2018) [72]	A prospective observational study aimed to improve the predicted ability in COPD progression	Results showed that machine learning methods improved the prediction accuracy of COPD progression	Not applicable	Not relevant to COPD phenotyping
Fang et al. (2018) [73]	A cross-sectional study aimed to estimate the COPD prevalence in China	Data showed that the estimated overall prevalence of COPD in China in 2014-15 was 13.6%	Not applicable	Not relevant to COPD phenotyping
Koo et al. (2018) [74]	A cross-sectional study aimed	Data showed that small		Not relevant to COPD

	to determine whether destruction of the terminal and transitional bronchioles occurs before, or in parallel with, emphysematous tissue destruction	airways disease is a pathological feature in mild and moderate COPD	Not applicable	phenotyping
Liang et al. (2018) [75]	A simulation study aimed to develop a novel variable reduction method for joint analysis of multiple phenotypes in association studies	Results showed that this novel method can be used in analyzing a whole-genome genotyping data	Not applicable	Not relevant with the machine learning methods under study
Raherison et al. (2018) [46]	A prospective observational study aiming to determine the association between specific comorbidities and COPD severity.	Cluster analysis identified five phenotypes of comorbidities	1) included cardiac profile, 2) included less comorbidities, 3) included metabolic syndrome, apnea and anxiety-depression, 4) included denutrition and osteoporosis, 5) included bronchiectasis	COPD phenotypes were not validated with clinical meaningful outcomes
Kilk et al. (2018) [76]	A pilot study aiming to characterize patients with COPD, based on the metabolomic profiling of peripheral blood and exhaled breath condensate (EBC) within the context of defined clinical and demographic variables.	Cluster analysis did not reveal a clinical-metabolomic stratification superior to the strata set by the GOLD consensus.	Not applicable	COPD phenotypes were not validated with clinical meaningful outcomes
de Vries et al. (2018) [47]	A multi-centre cross-sectional study to capture clinical/inflammatory phenotypes in patients with chronic airway disease using an electronic nose (eNose) in a training and validation set	Cluster analysis identified five combined asthma and COPD phenotypes	1) Asthma and COPD: predominantly females, high BMI, high symptom scores, low FeNO, no inflammation measured in blood, 2) Asthma and COPD: predominantly males, high circulating eosinophil counts, high FeNO, low use of oral	COPD phenotypes were not validated with clinical meaningful outcomes

			corticosteroids, 3) Asthma and COPD: predominately non-Caucasian, poor lung function, eosinophil blood counts of 0.45±1.3×109 cells·L-1, lowest exacerbation rate in the past 3 months, no OCS use, low use of ICS, 4) Asthma and COPD: predominantly atopic, high circulating neutrophil blood counts, highest number of exacerbations per person in the past 3 months, 5) fewer COPD patients, best postbronchodilator FEV1, relatively low exacerbation rate per person in the past 3 months	
Le Rouzic et al. (2018) [77]	A prospective observational study aimed to confirm the existence of the frequent exacerbator phenotype	Data confirmed the existence of the frequent exacerbator and the threshold to define this phenotype	Not applicable	COPD phenotypes were not validated with clinical meaningful outcomes
Vazquez Guillamet et al. (2018) [32]	A retrospective observational study aimed to identify COPD phenotypes from electronic medical records	Cluster analysis identified nine COPD phenotypes	1) depression–chronic obstructive pulmonary disease, 2) coronary artery disease–chronic obstructive pulmonary disease, 3) cerebrovascular disease–chronic obstructive pulmonary disease, 4) malignancy–chronic obstructive	COPD phenotypes were not validated with clinical meaningful outcomes

			pulmonary disease, 5) advanced malignancy–chronic obstructive pulmonary disease, 6) diabetes mellitus–chronic kidney disease–chronic obstructive pulmonary disease, 7) young age–few comorbidities–high readmission rates–chronic obstructive pulmonary disease, 8) atopy–chronic obstructive pulmonary disease, 9) advanced disease–chronic obstructive pulmonary disease	
Hall et al. (2018) [78]	An observational prospective study aimed to determine the extent to which multimorbidity is associated with long-term survival following acute myocardial infarction (AMI)	Three multimorbidity phenotype clusters that were significantly associated with loss in life expectancy were identified and should be a concomitant treatment target to improve cardiovascular outcomes.	Not applicable	Not relevant to COPD phenotyping
Das et al. (2018) [79]	A review of machine learning methods in the diagnosis of COPD	The application of artificial intelligence has produced promising results in the diagnosis of COPD	Not applicable	Not relevant to COPD phenotyping
Merchant et al. (2018) [80]	A prospective observational study aimed to assess the impact of digital intervention on asthma health resource utilization	Results showed that digital health interventions can be incorporated into routine clinical practice, and their use may contribute to improved outcomes including reduced healthcare utilization	Not applicable	Not relevant to COPD phenotyping

Radin et al. (2017) [31]	A cross-sectional study aimed to identify novel COPD phenotypes based on computed tomography (CT) densitometry	Cluster analysis showed the CT densitometry identified two distinct phenotypes of COPD	Cluster 1 has subjects with decreased FEV1, FEV/FVC, FEF at 25-75% of FVC and BMI and increased residual volume and total lung capacity compared to cluster 2	The derived phenotypes were not validated with clinical meaningful outcomes
Christenson et al. (2017) [81]	A randomized placebo-controlled clinical trial aimed to explore airway epithelial mucin gene expression heterogeneity in COPD	Cluster analysis identified that 2 COPD subgroups in which either MUC5AC or MUC5B gene expression is elevated. These subgroups are associated with specific inflammatory patterns	2 COPD subgroups in which either MUC5AC or MUC5B gene expression is elevated	The derived phenotypes were not validated with clinical meaningful outcomes
Kästle et al (2017) [82]	A genetic study aimed to identify specific miRNAs implicated in controlling Th17 differentiation	Results showed evidence of miRNAs involvement in controlling the differentiation and function of T helper cells, offering useful tools to study and modify Th17-mediated inflammation.	Not applicable	Not relevant with the machine learning methods under study
Fouda et al (2017) [83]	A prospective observational study on the association between osteoporosis and emphysema in a model that includes these potentially confounding factors	Results showed that emphysematous phenotype is not a risk factor for osteoporosis independently of BMI, FEV1, and PaO2.	Not applicable	Not relevant with the machine learning methods under study
Chalmers JD (2017) [84]	A review on bronchiectasis characterization	Key developments in the bronchiectasis field include the establishment of international disease registries and characterization of disease phenotypes using cluster analysis and biological data.	Not applicable	Not relevant to COPD phenotyping and machine learning methods under study
Fingleton et al. (2017)	A cross-sectional	Cluster analysis identified	1) severe late-onset	COPD phenotypes

[48]	observational study aiming to compare the phenotypes of airways disease in two separate populations (China and New Zealand)	five COPD phenotypes that were similar in both populations	asthma/COPD overlap group, 2) moderately severe early-onset asthma/COPD overlap group, 3) moderate to severe asthma group with type 2 predominant disease, 4 and 5) minimal airflow obstruction, differentiated by age of onset.	were not validated with clinical meaningful outcomes
Zarei et al. (2017) [35]	A randomized placebo-controlled trial aimed to identify COPD phenotypes using proteomic data	Cluster analysis identified three COPD phenotypes	The third cluster had less emphysema and worse disease-related quality of life, despite similar levels of lung function impairment than the other two groups	COPD phenotypes were not validated with clinical meaningful outcomes
Hirai et al. (2017) [85]	A prospective observational study aimed to clarify the discriminant factors for assigning the asthma-COPD overlap phenotype	Data showed that the asthma-COPD overlap phenotype was characterized by peripheral blood eosinophilia and higher levels of IgE despite the Th2-low endotype.	peripheral blood eosinophilia and higher levels of IgE despite the Th2-low endotype	COPD phenotypes were not validated with clinical meaningful outcomes
Lee et al. (2017) [49]	A national survey aimed to identify subtypes in patients with mild-to-moderate airflow limitation and to appreciate their clinical and socioeconomic implications	Cluster analysis identified five phenotypes with different level of health care utilization	1) near-normal: oldest mean age, highest FEV1, 2) asthmatic: youngest, lowest prescription rate,despite the highest proportion of self-reported wheezing, 3) chronic obstructive pulmonary disease (COPD): male predominant and all current or ex-smokers, high prescription rate of respiratory medicine, 4)	COPD phenotypes were not validated with clinical meaningful outcomes

			asthmatic-overlap: high prescription rate of respiratory medicine, 5) COPD-overlap: male predominant and all current or ex-smokers, high prescription rate of respiratory medicine.	
Haldar et al (2017) [86]	A genetic study aimed to assess whether the balance between the two dominant bacterial groups (Gammaproteobacteria (G) and Firmicutes (F)) in COPD sputum samples might reveal a subgroup with a bacterial community structure change at exacerbation that was restored to baseline on recovery and potentially reflects effective antibiotic treatment.	Results showed that the G:F ratio at exacerbation can be determined on a timescale compatible with decisions regarding clinical management	Not applicable	Not relevant to COPD phenotyping
Çolak et al. (2017) [87]	A prospective observational study aimed to investigate the prognosis of individuals with asymptomatic and symptomatic, undiagnosed COPD in the general population in Denmark.	Individuals with undiagnosed, symptomatic COPD had an increased risk of exacerbations, pneumonia, and death. Individuals with undiagnosed, asymptomatic COPD had an increased risk of exacerbations and pneumonia.	Not applicable	Not relevant to machine learning methods under study
Maddocks et al. (2016) [88]	A randomized placebo-controlled trial aimed to assess the effectiveness of neuromuscular electrical stimulation (NMES) as a home-based exercise therapy	Data showed that NMES improves functional exercise capacity in patients with severe COPD by enhancing quadriceps muscle mass and function.	Not applicable	Not relevant to COPD phenotyping
Lange et al. (2016) [89]	A prospective observational	Data showed that the		Not relevant to COPD

	study aimed to investigate the long-term prognosis of individuals with different types of chronic airway disease and asthma-COPD overlap	prognosis of individuals with asthma-COPD overlap is poor and seems to be affected by the age of recognition of asthma, being worst in those with late asthma onset (after 40 years of age)	Not applicable	phenotyping
Papi et al (2016) [24]	A prospective observational study aimed to define COPD phenotypes and identify biomarkers and/or genetic parameters that help to predict disease progression	The study highlights some of the progress in phenotyping the heterogeneity of the disease that have been made thanks to the analyses of this longitudinal study	Not applicable	Not relevant to machine learning methods under study
Ning et al. (2016) [40]	A cross-sectional analysis aimed to identify distinct COPD phenotypes	Cluster analysis identified four phenotypes	1) COPD patients with moderate to severe airflow limitation, 2) asthma and COPD patients with heavy smoking, airflow limitation and increased airways reversibility, 3) patients having less smoking and normal pulmonary function with wheezing but no chronic cough, 4) chronic bronchitis patients with normal pulmonary function and chronic cough	COPD phenotypes were not validated with clinical meaningful outcomes
Rootmensen et al. (2016) [41]	A cross-sectional study aimed to identify COPD phenotypes in an outpatient population	Cluster analysis identified four COPD phenotypes	1) patients with a history of extensive cigarette smoking, airway obstruction without signs of emphysema, 2) patients with features of the emphysematous	COPD phenotypes were not validated with clinical meaningful outcomes

			type of COPD, 3) patients with characteristics of allergic asthma, 4) patients with features suggesting an overlap syndrome of atopic asthma and COPD	
Sekiya et al. (2016) [50]	A prospective observational study aimed to examine the clinical characteristics and heterogeneity of patients with severe or life-threatening asthma exacerbation.	Cluster analysis identified five distinct asthma phenotypes	1) younger-onset asthma with severe symptoms at baseline, including limitation of activities, a higher frequency of treatment with oral corticosteroids and short-acting beta-agonists, and a higher frequency of asthma hospitalizations in the past year, 2) predominantly composed of elderly females, with the highest frequency of comorbid, chronic hyperplastic rhinosinusitis/nasal polyposis, and a long disease duration, 3) allergic asthma without inhaled corticosteroid use at baseline. Patients in this cluster had a higher frequency of atopy, including allergic rhinitis and furred pet hypersensitivity, and a better prognosis during hospitalization compared with the other	Not relevant to COPD phenotyping; not validated with clinical meaningful outcomes

			clusters, 4) elderly males with concomitant chronic obstructive pulmonary disease (COPD), 5) very mild symptoms at baseline according to the patient questionnaires, 41% had previously been hospitalized for asthma	
Sato et al. (2016) [90]	A retrospective study aiming to identify phenotypes of patients with idiopathic interstitial pneumonia (IIP) with pulmonary emphysema (PE)	Cluster analysis identified three phenotypes; idiopathic pulmonary fibrosis (IPF) with PE is a distinct phenotype with poor prognosis	Not applicable	Not relevant to COPD phenotyping
Morélot-Panzini et al. (2016) [91]	An observational prospective study testing the Multidimensional Dyspnea Profile (MDP) in COPD patients	The MDP can identify an affective/emotional dimension of dyspnea and contribute to phenotypic description of patients	Not applicable	COPD phenotypes were not validated with clinical meaningful outcomes
Roche et al. (2016) [92]	A cross-sectional study investigating the genetic variability of COPD and obstructive sleep apnea patients	The study identified genetic variants mapping to hypoxia response elements	Not applicable	Not relevant to COPD phenotyping
Martínez-García et al. (2016) [93]	An observational cohort study aimed to identify phenotypes for non-cystic fibrosis bronchiectasis	Using cluster analysis, it was possible to identify distinct phenotypes	Not applicable	Not relevant to COPD phenotyping
Batista-Navarro et al. (2016) [94]	A cross-sectional qualitative study that compared a manual performing task of COPD phenotype curation to that of a text-mining algorithm	Text-mining algorithms were more efficient in facilitating the curation of COPD phenotypes	Not applicable	Not relevant to methods under study; not validated with clinical outcomes
Labuzzetta et al. (2016) [95]	A genetic cross-sectional study that uses machine learning methods to predict COPD phenotypes	Machine learning methods showed that isoform expression data have high accuracy in predicting phenotypes	Not applicable	Predicted phenotypes not validated with clinical outcomes

Kaluarachchi et al. (2016) [96]	A case-control study aimed to determine perturbed biochemical functions associated with tobacco smoking	Results showed that combining multiplatform metabolic phenotyping with knowledge-based mapping gives mechanistic insights into disease development	Not applicable	Not relevant to COPD phenotyping
Obeidat et al. (2015) [97]	A genome-wide association study aimed to investigate molecular mechanisms underlying variations in lung function	The system genetics approach identified lung tissue genes driving the variation in lung function and susceptibility to COPD	Not applicable	Not relevant to COPD phenotyping
Kim et al. (2015) [98]	A cross-sectional study aimed to identify novel lung disease phenotypes using multi-omics data	Cluster analysis identified subclusters with distinct clinical and biomolecular characteristics	Not applicable	Not relevant to COPD phenotyping; not validated with clinical meaningful outcomes
Hübenthal et al. (2015) [99]	A case-control study that used genetic profiling and machine learning methods to accurately predict inflammatory diseases	The proposed miRNA signature is of relevance for the etiology of inflammatory bowel disease (IBD)	Not applicable	Not relevant to COPD phenotyping
Lee et al. (2014) [100]	An observational genetic study aimed to investigate the clinical and genetic heterogeneity in subjects with mild airflow limitation in spirometry grade 1 defined by the Global Initiative for COPD	Results showed that GOLD 1 subjects show substantial clinical heterogeneity, which is at least partially related to genetic heterogeneity.	Not applicable	The derived phenotypes were not validated with clinical meaningful outcomes
Uzun et al. (2014) [101]	A randomized placebo-controlled trial aimed to investigate whether patients with COPD who had received treatment for three or more exacerbations in the previous year would have a decrease in exacerbation rate when maintenance treatment with azithromycin was added to standard care	Data showed that maintenance treatment with azithromycin significantly decreased the exacerbation rate compared with placebo	Not applicable	Not relevant to COPD phenotyping
Brightling et al. (2014) [102]	A randomized placebo-controlled trial aimed to	Results showed that compared with placebo,	Not applicable	Not relevant to COPD phenotyping

	establish whether benralizumab reduces acute exacerbations of COPD in patients with eosinophilia and COPD	benralizumab did not reduce the rate of acute exacerbations of COPD		
Kon et al. (2014) [103]	Three prospective observational studies aimed to assess the he minimum clinically important difference (MCID) for the COPD Assessment Test (CAT) in patients with COPD	The most reliable estimate of the minimum important difference of the CAT is 2 points	Not applicable	Not relevant to COPD phenotyping
Köhnlein et al. (2014) [104]	A prospective randomized controlled clinical trial aimed to investigate the effect of long-term non-invasive positive pressure ventilation (NPPV), targeted to markedly reduce hypercapnia, on survival in patients with advanced, stable hypercapnic COPD	Results showed that the addition of long-term NPPV to standard treatment improves survival of patients with hypercapnic, stable COPD when NPPV is targeted to greatly reduce hypercapnia.	Not applicable	Not relevant to COPD phenotyping
Jones et al. (2014) [105]	A retrospective study aimed to investigate patterns of health-care use and comorbidities present in patients in the period before diagnosis of chronic obstructive pulmonary disease (COPD)	Data showed that opportunities to diagnose COPD at an earlier stage are being missed, and could be improved by case-finding in patients with lower respiratory tract symptoms and concordant long-term comorbidities.	Not applicable	Not relevant to COPD phenotyping
Zheng et al. (2014) [106]	A randomized placebo-controlled trial aimed to assess whether N-acetylcysteine could reduce the rate of exacerbations in patients with COPD	Data showed that in Chinese patients with moderate-to-severe COPD, long-term use of N-acetylcysteine 600 mg twice daily can prevent exacerbations, especially in disease of moderate severity.	Not applicable	Not relevant to COPD phenotyping
Corhay et al. (2014)	A cross-sectional study aimed	Cluster analysis can help		COPD phenotypes

[107]	to summarize the current data available about the phenotypes of this disease	to identify more precise definition of COPD phenotypes	Not applicable	were not validated with clinical meaningful outcomes
Moore et al. (2014) [108]	A cross-sectional study aiming to understand the interactions between inflammation and clinical asthma subphenotypes	Cluster analysis identified four phenotypes associated with asthma severity	Not applicable	Not relevant to COPD phenotyping
Qiao et al. (2014) [109]	A simulation study investigating the association between genetic loci and complex phenotypes	Cluster analysis can be useful in genome sequencing studies for pairing genomic regions with complex phenotypes	Not applicable	Not relevant to COPD phenotyping
DiSantostefano et al. (2014) [110]	Baseline data of two clinical trials were used to identify risk groups for pneumonia	Cluster analysis can identified distinct patient groups at risk of pneumonia	Not applicable	Not relevant to COPD phenotyping
Vogelmeier et al. (2013) [111]	A randomized parallel group trial aimed to compare the efficacy, safety, and tolerability of QVA149 versus salmeterol-fluticasone (SFC) over 26 weeks in patients with moderate-to-severe COPD	Results suggested the potential of dual bronchodilation as a treatment option for non-exacerbating symptomatic COPD patients	Not applicable	Not relevant to COPD phenotyping
Franciosi et al. (2013) [112]	Four clinical trials aimed to assess the efficacy and safety of a novel inhaled dual phosphodiesterase 3 (PDE3) and PDE4 inhibitor, RPL554 for its ability to act as a bronchodilator and anti-inflammatory drug	Data showed that inhaled RPL554 is an effective and well tolerated bronchodilator, bronchoprotector, and anti-inflammatory drug	Not applicable	Not relevant to COPD phenotyping
Decramer et al. (2013) [113]	A randomized parallel group study aimed to compare the efficacy and safety of indacaterol and tiotropium in patients with COPD	Data showed that Indacaterol and tiotropium provided clinically relevant improvements in lung function with comparable safety profiles.	Not applicable	Not relevant to COPD phenotyping
Dransfield et al. (2013) [114]	Two parallel group randomized controlled trials aimed to investigate whether fluticasone	Results showed that addition of fluticasone furoate to vilanterol was	Not applicable	Not relevant to COPD phenotyping

	furoate and vilanterol would prevent more exacerbations than would vilanterol alone	associated with a decreased rate of moderate and severe exacerbations of COPD in patients with a history of exacerbation, but was also associated with an increased pneumonia risk		
Wedzicha et al. (2013) [115]	A randomized parallel-group study aimed to evaluate the effect of dual, longacting inhaled bronchodilator treatment on exacerbations in patients with severe and very severe chronic obstructive pulmonary disease (COPD)	Results suggested potential of dual bronchodilation as a treatment option for patients with severe and very severe COPD.	Not applicable	Not relevant to COPD phenotyping
Rabe et al. (2013) [116]	A randomized parallel-group study aimed to establish whether ADRB2 polymorphisms differentially affected COPD exacerbation outcomes in response to tiotropium versus salmeterol.	Data showed limited evidence for the use of ADRB2 polymorphisms for predicting LABA treatment response	Not applicable	Not relevant to COPD phenotyping
Siedlinski et al. (2013) [117]	A case-control study aimed to estimate direct and indirect effects of genetic loci on COPD development using mediation analysis	This study confirms the existence of direct effects of the AGPHD1/CHRNA3, IREB2, FAM13A and HHIP loci on COPD development.	Not applicable	Not relevant to COPD phenotyping and machine learning methods under study
Gouzi et al. (2013) [118]	A cross-sectional study aimed to test whether muscle fiber atrophy and increased oxidative stress constitute the attributes of validated COPD phenotypes	Data showed that demonstrates that the muscle heterogeneity is the translation of different phenotypes of the disease.	Not applicable	COPD phenotypes were not validated with clinical meaningful outcomes
Fens et al. (2013) [42]	A cross-sectional study aimed to identify subphenotypes of COPD in a community-based population of heavy (ex-) smokers	Cluster analysis identified four COPD phenotypes	1) mild COPD, limited symptoms and good quality of life, 2) low lung function, combined emphysema and chronic	COPD phenotypes were not validated with clinical meaningful outcomes

			bronchitis and a distinct breath molecular profile, 3) emphysema predominant COPD with preserved lung function, 4) highly symptomatic COPD with mildly impaired lung function.	
Shaykhiev et al. (2013) [119]	A genetic study investigating the association between CXCL14 gene, cancer and COPD	Data showed that smoking-induced gene expression is a potential link between smoking-associated airway epithelial injury, COPD, and lung cancer.	Not applicable	Not relevant to COPD phenotyping
Carolan et al. (2013) [120]	A review that discusses advances in describing phenotypic variability in asthma and COPD	The authors suggest that better understanding of the heterogeneity of the disease through phenotyping will improve care and reduce potential adverse effects from unnecessary therapies	Not applicable	Not relevant to methods under study.i.e. a review - not original research study
Basagaña et al. (2013) [39]	In this article the authors developed a framework of applying imputation to missing values of a cluster analysis	The proposed framework deals with uncertainty in definine the number of clusters, the variable selection and allocation of patients to clusters	Not applicable	Not relevant to the studies under review
Toraldo et al. (2012) [121]	A review article that discusses and refines the concept of desaturator phenotypes in COPD with pulmonary hypertension (PH)	Cluster analysis can identify a pattern of phenotypic markers that could be used as a framework for future diagnosis and research	Not applicable	Not relevant to COPD phenotyping and machine learning methods under study
Travers et al. (2012) [122]	In a letter to the editors the authors discuss the possibility of re-examining the classification of airways disease to identify disease subgroups that may respond to	The authors conclude that classification analysis can be used to derive allocation rules that allow disease groups identified through cluster analysis to	Not applicable	Not relevant to machine learning methods under study

	treatments in different ways.	be prospectively identified in the real world. This will enable trials to test interventions in putative phenotypes, a necessary step towards personalised medicine for airways disease.		
Toraldo et al. (2011) [123]	A cross-sectional study aimed to discuss and refine the concept of phenotyping desaturators in COPD and shows a possible pattern which could be used as a framework for future research.	The study suggests that COPD phenotyping can facilitate our understanding and management of COPD	Not applicable	Not relevant to machine learning methods under study
Bafadhel et al. (2011) [36]	A cross-sectional study aimed to study the application of CT imaging in the multidimensional approach to phenotyping patients with COPD	Cluster analysis identified three clusters, two of which were emphysema predominant and the third characterized by a heterogeneous combination of emphysema and bronchiectasis	1) emphysema (EM) predominant, 2) bronchiectasis (BE) predominant, 3) heterogeneous combination of EM and BE	The derived phenotypes were not validated with clinical meaningful outcomes
Bafadhel et al. (2011) [43]	A prospective observational study aimed to investigate biomarker expression in COPD exacerbations to identify biologic clusters and determine biomarkers that recognize clinical COPD exacerbation phenotypes	Cluster analysis identified four distinct biologic exacerbation clusters	1) bacterial-predominant, 2) viral-predominant, 3) eosinophilic--predominant, 4) limited changes in the inflammatory profile	COPD phenotypes were not validated with clinical meaningful outcomes
Fingleton et al. (2011) [124]	In a letter to the editors the authors discuss the tailoring of treatment regiments to patients with different COPD phenotypes	The author acknowledge the challenge to determine distinct phenotypes and suggest that if these phenotypes are validated with response to treatment then can be potentially used to target treatments	Not applicable	Not relevant to methods under study, i.e. a review - not original research study

		specifically to patients		
Shirtcliffe et al. (2011) [125]	This review aimed to a better understanding of the distinct disorders of airways disease with the potential to inform on underlying mechanisms, risk factors, natural history, monitoring and treatment.	The authors conclude that by further defining the distinct phenotypes that make up the syndromes of asthma and COPD could lead to treatments specifically targeted for defined phenotypic groups.	Not applicable	Not relevant to methods under study, i.e. a review - not original research study
Sharma et al. (2010) [126]	A study used data from two clinical trials aimed to identify subject clusters in one study and replicate the findings in the second study	Cluster analysis identified three subjects clusters in one study that were replicated in the second study	Not applicable	The derived phenotypes were not validated with clinical meaningful outcomes
Jo et al. (2010) [127]	A cross-sectional observational study aimed to classify the phenotypes in elderly subjects with obstructive lung disease (OLD)	Cluster analysis identified three phenotypes in elderly patients with OLD	Not applicable	The derived phenotypes were not validated with clinical meaningful outcomes
Cho et al. (2010) [44]	An observational genetic study aimed to identify subtypes of severe emphysema	Cluster analysis identified four phenotypes in a group of sever emphysema patients	1) emphysema predominant, 2) bronchodilator responsive, with higher FEV1, 3) discordant, with a lower FEV1 despite less severe emphysema and lower airway wall thickness, 4) airway predominant.	The derived phenotypes were not validated with clinical meaningful outcomes
Sobradillo et al. (2010) [128]	In this article the authors review the knowledge in the topic of COPD phenotypes		Not applicable	Not relevant to the purpose of the review under study
Weatherall et al. (2010) [129]	In this article the authors discuss the advantages and disadvantages of cluster analysis to characterize different types of airways disorders	The author conclude that cluster analysis can help to better understanding the true patterns of airway disorders and could lead to different pharmacological treatments and other interventions directed at	Not applicable	Not relevant to machine learning methods under study

		specific phenotypic group		
Paoletti et al. (2009) [130]	A cross-sectional study aimed to assess the presence of hidden structures in data corresponding to the different COPD phenotypes observed in clinical practice	Data showed that using cluster analysis can identify phenotypes for understanding the results of pharmacologic trials; clinician's approach to patient treatment and COPD natural history.	Not applicable	The derived phenotypes were not validated with clinical meaningful outcomes
Weatherall et al. (2009) [51]	A cross-sectional study aimed to explore clinical phenotypes in a community population with airways disease	Cluster analysis identified five distinct phenotypes of airflow obstruction	1) severe and markedly variable airflow obstruction with features of atopic asthma, chronic bronchitis and emphysema, 2) features of emphysema alone, 3) atopic asthma with eosinophilic airways inflammation, 4) mild airflow obstruction without other dominant phenotypic features, 5) chronic bronchitis in nonsmokers	The derived phenotypes were not validated with clinical meaningful outcomes
Pistolesi et al. (2008) [131]	A cross-sectional study aimed to ascertain whether COPD phenotypes reflecting different mechanisms of airflow limitation could be clinically identified	Results showed that patients with COPD can be assigned a clinical phenotype reflecting the prevalent mechanism of airflow limitation	Not applicable	Not relevant to COPD phenotyping and machine learning methods under study
Patel et al. (2008) [132]	An observational study aiming to assess the association between airway wall thickening and emphysema at the severity of COPD	Airway wall thickening and emphysema make independent contributions to airflow obstruction in COPD.	Not applicable	Not relevant with machine learning methods under study
Kodavanti et al. (2006) [133]	An animal study investigating whether spontaneously hypertensive (SH) rats may offer a better model of experimental bronchitis and	Data showed that sulfur dioxide (SO2) exposure SH rats may yield a relevant experimental model of bronchitis	Not applicable	Not relevant to COPD phenotyping

	subsequent COPD phenotypes			
Wardlaw et al. (2005) [134]	An article that discusses the use of a new taxonomy for mutli-dimensional phenotyping	The authors suggest that development of this taxonomy will require a much more complete and sophisticated correlation of the many variables that uses complex statistical tools such as cluster analysis	Not applicable	Not relevant machine learning methods under study
Hackett et al. (2003) [135]	A genetic study investigating the association between antioxidant-related genes and smoking-induced chronic bronchitis	Data showed that antioxidant-related genes may be useful genetic markers in assessing susceptibility to smoking-induced chronic bronchitis	Not applicable	Not relevant to COPD phenotyping

Table 4. Data characteristics and methods for the identification of COPD phenotypes in the reviewed studies

Study	Data used in the clustering analysis	Data reduction and clustering methods
Yoon et al. (2019) [9]	Age, BMI, smoking status, history of asthma, COPD assessment test (CAT) score, pre-bronchodilator FEV1 % predicted, diffusing capacity of carbon monoxide % predicted	K-means
Pikoula et al. (2019) [6]	BMI, smoking status, atopy, GINA1 classification, eosinophilia, comorbidities	Multiple correspondence analysis (MCA), k-means, and hierarchical clustering
Kim et al. (2018) [10]	BMI, Charlson comorbidity index, SGRQ2 total score, FEV1	Factor analysis and hierarchical clustering
Kim et al. (2017) [11]	Clinical, physiological and imaging data	PCA and hierarchical cluster analysis
Burgel et. al. (2017) [8]	Age, BMI, FEV1 % predicted, mMRC3 dyspnea scale, exacerbation in the past 12 months, comorbidities	Factor analysis for mixed data (FAMD) and hierarchical clustering
Peters et al. (2017) [14]	FEV1 % predicted, BMI, exercise capacity, subjective symptoms, fatigue, quality of life	Hierarchical and discriminant cluster analysis
Chubachi et al. (2016) [17]	Comorbidity data (e.g., cardiovascular diseases and diabetes)	Hierarchical cluster analysis
Fingleton et al. (2015) [19]	Respiratory history and comorbidities, lung function, reversibility testing, biomarkers, disease control and health status	Hierarchical cluster analysis
Chen et al (2014) [16]	Age, lung function (FEV1 % predicted), BMI, history of severe exacerbations, mMRC, SpO2, Charlson Index	PCA, hierarchical, and k-means clustering
Castaldi et al. (2014) [7]	Demographic and clinical characteristics, spirometry, genome-wide SNP genotyping data, inspiratory and expiratory CT scans	Factor analysis and k-means clustering
Altenburg et al. (2012) [18]	Age, BMI, quadriceps force, body plethysmography, exercise testing	K-means cluster analysis
Burgel et al. (2010) [12]	Age, symptoms, spirometry, BMI, exacerbations, health and psychological status	PCA and hierarchical cluster analysis
Burgel et al. (2012) [13]	Age, symptoms, health status, body plethysmography, DLCO4, CT scan, comorbidities	PCA and hierarchical cluster analysis
Garcia-Aymerich et al. (2011) [15]	Symptoms, health status, body composition, plethysmography, CT scan, saliva and serum, exercise testing	K-means cluster analysis

[1]GINA: Global Initiative for Asthma; [2]SGRQ: St George's Respiratory Questionnaire; [3]mMRC: Modified Medical Research Council;

[4] DL_{CO}: Diffusing capacity of the lungs for carbon monoxide

Table 5. Best practices recommended for the identification of clinically validated COPD phenotypes using clustering analysis

Prospective longitudinal data	**External validation**	**Large samples**	**Handling of missing data**	**Choice of variables and cluster analysis**
Use longitudinal prospective data over a long period of time from a large database (e.g., CALIBER, UK Biobank)	Cross-validation with different databases from multiple settings (in different parts of the world), and validation against clinically meaningful endpoints (e.g., exacerbations, response to therapy, mortality)	Use large samples, ideally with more than 1,000 patients	Multiple imputation methods and sensitivity analysis	Through a combination of expert opinions, evidence-based data and literature reviews, data reduction methods, and cluster analysis

Figure 1. PRISMA diagram for the systematic review