CHEST

Machine Learning Characterization of COPD Subtypes

Insights From the COPDGene Study

Peter J. Castaldi, MD; Adel Boueiz, MD; Jeong Yun, MD; Raul San Jose Estepar, PhD; James C. Ross, PhD; George Washko, MD; Michael H. Cho, MD; Craig P. Hersh, MD; Gregory L. Kinney, PhD; Kendra A. Young, PhD; Elizabeth A. Regan, MD; David A. Lynch, MD; Gerald J. Criner, MD; Jennifer G. Dy, PhD; Stephen I. Rennard, MD; Richard Casaburi, MD; Barry J. Make, MD; James Crapo, MD; Edwin K. Silverman, MD, PhD; and John E. Hokanson, PhD; for the COPDGene Investigators[*]

COPD is a heterogeneous syndrome. Many COPD subtypes have been proposed, but there is not yet consensus on how many COPD subtypes there are and how they should be defined. The COPD Genetic Epidemiology Study (COPDGene), which has generated 10-year longitudinal chest imaging, spirometry, and molecular data, is a rich resource for relating COPD phenotypes to underlying genetic and molecular mechanisms. In this article, we place COPDGene clustering studies in context with other highly cited COPD clustering studies, and summarize the main COPD subtype findings from COPDGene. First, most manifestations of COPD occur along a continuum, which explains why continuous aspects of COPD or disease axes may be more accurate and reproducible than subtypes identified through clustering methods. Second, continuous COPD-related measures can be used to create subgroups through the use of predictive models to define cut-points, and we review COPDGene research on blood eosinophil count thresholds as a specific example. Third, COPD phenotypes identified or prioritized through machine learning methods have led to novel biological discoveries, including novel emphysema genetic risk variants and systemic inflammatory subtypes of COPD. Fourth, trajectory-based COPD subtyping captures differences in the longitudinal evolution of COPD, addressing a major limitation of clustering analyses that are confounded by disease severity. Ongoing longitudinal characterization of subjects in COPDGene will provide useful insights about the relationship between lung imaging parameters, molecular markers, and COPD progression that will enable the identification of subtypes based on underlying disease processes and distinct patterns of disease progression, with the potential to improve the clinical relevance and reproducibility of COPD subtypes.

CHEST 2020; 157(5):1147-1157

KEY WORDS: COPD; emphysema; machine learning

ABBREVIATIONS: GOLD = Global Initiative for Chronic Obstructive Lung Disease; ICGC = International COPD Genetics Consortium; TGF-β = transforming growth factor-β

AFFILIATIONS: From the Channing Division of Network Medicine (Drs Castaldi, Boueiz, Yun, Cho, Hersh, and Silverman), General Medicine and Primary Care (Dr Castaldi), Pulmonary and Critical Care Medicine (Drs Boueiz, Yun, Washko, Cho, Hersh, and Silverman), and Applied Chest Imaging Laboratory (Drs Estepar, Washko, and Ross), Brigham and Women's Hospital, Harvard Medical School, Boston, MA; Department of Epidemiology (Drs Kinney, Young, and Hokanson), University of Colorado, Denver, Aurora, CO; National Jewish Health, Department of Medicine (Drs Regan, Make, and Crapo), Denver, CO; Department of Radiology (Dr Lynch), National Jewish Health, Denver, CO; Department of Thoracic Medicine and Surgery (Dr Criner), Lewis Katz School of Medicine at Temple University, Philadelphia, PA; Department of Electrical and Computer Engineering (Dr Dy), Northeastern University, Boston, MA; Pulmonary and Critical Care Medicine (Dr Rennard), University of Nebraska Medical Center, Omaha, NE; and the Rehabilitation Clinical Trials Center (Dr Casaburi), Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA.

*Collaborators from the COPDGene Investigators are listed in the Acknowledgments.

FUNDING/SUPPORT: The project described was supported by the National Heart, Lung, and Blood Institute [Awards U01 HL089897, U01 HL089856, and R01 HL124233]. The COPD Genetic Epidemiology Study (COPDGene) is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprising AstraZeneca, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, and Sunovion.

COPD has many different clinical presentations, and COPD can be viewed as an umbrella syndrome that encompasses many distinct diseases.[1] Despite recent efforts to expand the criteria for diagnosing and staging COPD, the definitions from expert panels[2] do not fully capture the clinical heterogeneity of the disease.

The COPD Genetic Epidemiology Study (COPDGene) has generated detailed, longitudinal clinical phenotyping and genomic data for thousands of smokers, and these data are a rich resource for understanding the clinical and molecular heterogeneity of COPD. Machine learning methods can be used to identify new subtypes of COPD, defined by using patterns of clinical and molecular markers. Dozens of articles using COPDGene data have addressed this question, but there has been no comprehensive review of these scientific contributions.

The current article reviews the most relevant subtyping articles from COPDGene according to the broad questions they address: (1) How can clustering methods be used to discover novel subtypes, and are these subtypes reproducible? (2) What other machine learning methods besides clustering can be used to study COPD heterogeneity? (3) How can cut-points be defined in a data-driven way to turn continuous COPD measures into subtypes? (4) How can machine learning on chest CT data improve our ability to characterize COPD heterogeneity? (5) Are there distinct trajectories of lung function over the life course that correspond to molecular subtypes of COPD? In addition to this review, the contributions of COPDGene to COPD imaging,[3] physiology,[4] clinical epidemiology,[5] genetics,[6] and biomarker discovery[7] have been covered in separate reviews.

The following sections provide a brief background on the study of COPD subtypes and the use of unsupervised machine learning methods for disease subtyping, and we summarize the most important published results in this area using COPDGene data.

Historical Perspective on COPD Subtypes

Clinicians and COPD researchers have long recognized that COPD encompasses multiple different disease processes. However, it has been difficult to precisely define the molecular underpinnings of the diverse phenotypic manifestations of COPD. As a result, the COPD field lacks the information required to develop a sufficiently detailed, comprehensive disease classification. The 1958 CIBA Symposium was a landmark event in COPD subtyping, and the summary of this symposium states that the lack of a precise COPD definition resulted in "confusion and misunderstanding between investigators working in different centers and in different branches of medicine" that limited the fundamental understanding of COPD.[8] The CIBA Symposium framework remains influential today, particularly with respect to: (1) pathologic classification of emphysema based on the anatomy of the secondary pulmonary lobule; (2) the differentiation of reversible (asthma-related) from irreversible (COPD-related) pulmonary obstruction; and (3) the identification of chronic bronchitis and emphysema as the two primary clinical phenotypes of COPD.

In subsequent work, Charles Fletcher and Benjamin Burrows expanded on the concept of the chronic bronchitis and emphysema-predominant subtypes of COPD by using a variety of clinical measurements to define type A (emphysema-predominant) and type B (bronchial) COPD subtypes.[9,10] Notably, this classification also included type X patients who did not meet criteria for either category, and although the authors provided general outlines for these subtypes, they concluded that "firm definitions of the syndromes would be premature" due to lack of understanding of the etiologic mechanisms of COPD. In subsequent years, multiple additional COPD subtypes were proposed, including the frequent exacerbator subtype,[11,12] asthma-COPD overlap,[13] and upper lobe-predominant emphysema.[14]

With the advent of larger datasets, machine learning methods were used for COPD subtype discovery,[15-20] and a selected list of such studies in included in Table 1. However, these clustering studies used different methods and variables, making it challenging to synthesize and interpret this literature.[21]

In addition, there are fundamentally different perspectives on whether COPD is best described by using distinct subgroups or rather multiple overlapping disease processes. The term COPD "subtypes" has two common uses. It refers broadly to the study of COPD heterogeneity, but in its more specific meaning it refers to distinct, nonoverlapping subgroups of subjects. The

CORRESPONDENCE TO: Peter J. Castaldi, MD, Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Ave, Boston, MA 02115; e-mail: repjc@channing.harvard.edu

DOI: https://doi.org/10.1016/j.chest.2019.11.039

TABLE 1] Selected Publications Using Machine Learning Methods to Identify Clusters or Disease Axes in COPD

Category	PMID	Year	No. of Subjects	No. of Clusters/Axes	Method
Clustering	18248806	2008	415	2 clusters	Fuzzy clustering
	19501190	2009	415	2 clusters	Multidimensional scaling and KHM clustering
	20233420	2010	308	4 clusters	K-means
	20075045	2010	322	4 clusters	Principal components analysis and hierarchical clustering
	21177668	2011	342	3 clusters	K-means
	22154126	2012	102	2 clusters	K-means
	23236428	2012	527	3 clusters	Principal components analysis and hierarchical clustering
	23392440	2013	213	5 clusters	Self-organizing maps
	23613569	2013	1,543	3 clusters	Tree-based clustering
	23536961	2013	157	4 clusters	Factor analysis and k-means
	24563194	2014	8,288	4 clusters	K-means
	25642832	2015	2,164	5 clusters	Factor analysis and random forests clustering
	26773458	2016	364	4 clusters	Network-based stratification
	28943279	2017	9,210	3 clusters	Random forests clustering
	29097431	2017	6,060	5 clusters	Hierarchical clustering
	28637835	2017	17,146	Multiple solutions	Random forests and k-medoids clustering
	29671603	2018	4,606	4 trajectories	Bayesian trajectory modeling
Disease axes	19480658	2009	127	4 disease axes	Principal components analysis
	29771274	2018	8,157	5 disease axes	Factor analysis
	31189730	2019	4,726	6 disease axes	Weighted logistic regression

term “endotype”[22] refers to underlying molecular processes that define subtypes, similar to the concept of T-helper type 2-mediated airway inflammation in asthma.[23] Unlike subtypes or endotypes, the term “treatable traits”[24] was proposed as an alternative to the concept of subtypes in which rigid subgroup boundaries were replaced by a more flexible characterization based on overlapping traits, such as bronchodilator responsiveness, airway wall thickening, and sputum eosinophilia. In the treatable traits paradigm, subjects with COPD can have many overlapping disease processes that may vary in severity, rather than being classified into one and only one subtype. A similar concept has also been proposed in diabetes.[25] Finally, the term “disease axis”[26] refers specifically to continuous measures that are composed of many contributing variables. Disease axes are produced by a specific class of machine learning methods called dimension reduction algorithms, and they were proposed as an alternative to clustering algorithms for COPD subtyping.

Challenges and Applications of Unsupervised Machine Learning Methods in COPD Subtyping

Machine learning refers to the design, development, and analysis of computational algorithms that automatically “learn” from experience (data) to achieve a specific task. In COPD, unsupervised learning algorithms have been used to discover novel subtypes by mining complex datasets. Two major classes of unsupervised machine learning algorithms are clustering and dimension reduction. Clustering algorithms such as k-means or hierarchical clustering seek to assign subjects into groups by some measure of similarity. In this sense, clustering methods simplify data along the subject dimension by compressing a large number of subjects into a smaller number of groups or clusters. When the data do not intrinsically have distinct clusters, the choice of cluster number can be arbitrary, and highly dataset- and method-dependent.

Dimension reduction methods simplify datasets along the variable dimension by combining measured variables into a smaller number of composite variables that contain as much of the original information as possible. Dimension reduction is most useful when there is strong correlation structure in a dataset, because much of the information can be "compressed" into a smaller number of composite variables, thereby reducing the dimension of the original dataset.

With the increasing availability of data-rich measurements such as CT images and genomic datasets in thousands of subjects with COPD, machine learning has the potential to discover novel connections between the physiologic manifestations of COPD and their underlying biological processes. However, the application of machine learning to COPD subtyping faces many challenges. Machine learning algorithms are complex and do not always produce results that are reliable or readily interpretable. Effective applications of machine learning often still rely on human expertise to extract the proper meaning from noisy variables and to evaluate between multiple possible outputs from the same algorithm. The current article illustrates the limitations of machine learning in COPD subtyping and some of the successes to date.

COPDGene Contributions to the Identification of COPD Subtypes and Disease Axes

COPDGene enrolled a total of 10,192 current and former smokers across the full spectrum of lung function at 21 different centers across the United States.[27] At baseline, 43% of subjects had normal spirometry findings, and 36% were in Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage 2, 3, or 4. Two-thirds of the subjects were non-Hispanic white, and one-third were African American. Forty-seven percent were women, and the average age of subjects at enrollment was 60 years. Nearly all study subjects underwent spirometry, questionnaire assessments, standardized inspiratory and expiratory chest CT imaging, and genome-wide genotyping. Five-year follow-up data were obtained for 6,758 subjects, and 10-year visits are currently being conducted. Figure 1 provides an overview of the number of subjects and data types currently available for each of the three COPDGene visits, and Figure 2 provides an overview of the major findings from machine learning analyses of COPD subtypes and disease axes in COPDGene data.

How Can Clustering Methods Be Used to Discover Novel Subtypes, and Are These Subtypes Reproducible?

To identify COPD subtypes using clinical variables, Castaldi et al[18] analyzed spirometric and imaging variables using k-means clustering to identify four clusters of phenotypically distinct subjects in COPDGene. These clusters were: (1) relatively resistant to smoking; (2) mild upper lobe emphysema-predominant; (3) airway-predominant COPD; and (4) severe airflow obstruction and emphysema. Although the average characteristics of the four clusters were distinct, when we visualized the clusters, there was little separability between the groups (Fig 3A), indicating that the subjects in COPDGene are distributed along a continuous spectrum of phenotypic variability, rather than forming clearly distinct clusters.

When genetic association testing was performed for these clusters, the severe obstruction/emphysema and the upper lobe-predominant groups exhibited a strong association with several known COPD-associated variants, whereas the airway-predominant groups had a much weaker pattern of genetic association. The observation of strong genetic associations to the mild upper lobe-predominant groups led to subsequent articles examining the genetic basis of apico-basal emphysema distribution. A genome-wide association study for emphysema distribution identified five genome-wide significant associations,[28] and subsequent cell-based functional studies identified an emphysema-associated functional variant altering the expression of *ACVR1B*, a signaling receptor in the transforming growth factor-β (TGF-β) superfamily.[29] In a separate clustering analysis focused specifically on measures of emphysema distribution, the upper lobe-predominant group was observed to have more rapid 5-year progression of emphysema in both unadjusted and multivariate adjusted analyses.[30]

To determine whether blood gene expression data can be used to stratify smokers according to systemic inflammation state, Chang et al[31] applied a network-based stratification method to gene expression data from subjects from COPDGene and Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) studies. This analysis identified reproducible gene expression signatures that distinguished four subtypes of smokers. These signatures distinguished subjects with moderate airflow obstruction from those without obstruction; in addition, the signatures of some subgroups were enriched for inflammatory pathways such as IL6-JAK-STAT signaling, in which the expression of this pathway was increased in the cluster with the lowest average FEV_1 relative to the cluster with the highest FEV_1. Other gene functional categories such as lymphocyte activation, wound healing, and protein catabolism were also associated with subtype signatures.

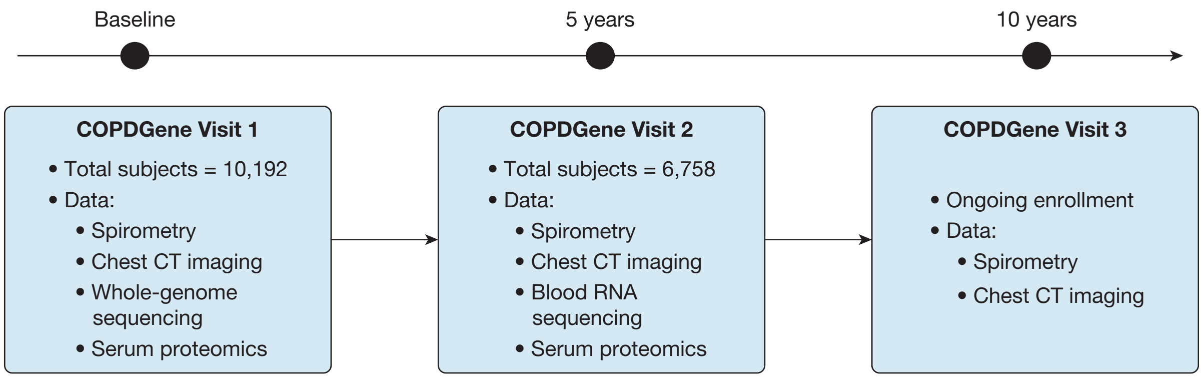

Figure 1 – *Overview of data gathered at the baseline, 5-year, and 10-year visits of the COPD Genetic Epidemiology Study (COPDGene).*

When we compared the overlap between the clustering assignments for 120 COPDGene subjects included in both the Castaldi et al[18] and Chang et al[31] articles, the clusterings were different (Table 2).

Because the studies by Castaldi et al[18] and Chang et al[31] used different input variables, it is not surprising that the clusters differed. However, for studies evaluating similar variables, one would expect that clustering studies across different cohorts would produce similar results. In fact, when comparing the average characteristics of clusters, the subtypes identified by Castaldi et al do show some similarity to those reported in other studies. In 342 subjects with COPD hospitalized for respiratory exacerbation,[16] three clusters were identified, two of which resembled the airway-predominant COPD and severe airflow and emphysema clusters. The upper lobe-predominant emphysema cluster was not identified in this study, which was expected because it did not include CT-quantified emphysema. In another study of 415 subjects with COPD recruited from outpatient clinics,[15] two clusters were identified that again resembled the airway-predominant COPD and severe airflow and emphysema groups. In the only systematic review conducted of COPD clustering studies, Pinto et al[21]

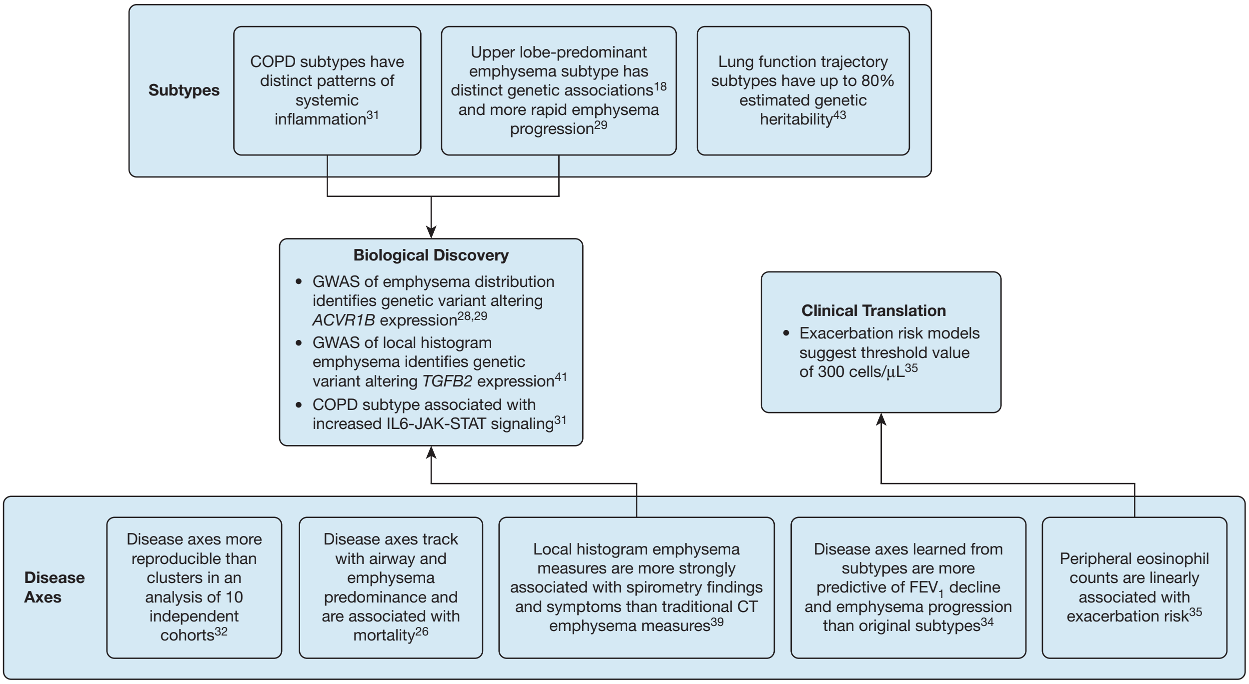
Subtypes
COPD subtypes have distinct patterns of systemic inflammation31
Upper lobe-predominant emphysema subtype has distinct genetic associations18 and more rapid emphysema progression29
Lung function trajectory subtypes have up to 80% estimated genetic heritability43
Biological Discovery
• GWAS of emphysema distribution identifies genetic variant altering ACVR1B expression28,29
• GWAS of local histogram emphysema identifies genetic variant altering TGFB2 expression41
• COPD subtype associated with increased IL6-JAK-STAT signaling31
Clinical Translation
• Exacerbation risk models suggest threshold value of 300 cells/µL35
Disease Axes
Disease axes more reproducible than clusters in an analysis of 10 independent cohorts32
Disease axes track with airway and emphysema predominance and are associated with mortality26
Local histogram emphysema measures are more strongly associated with spirometry findings and symptoms than traditional CT emphysema measures39
Disease axes learned from subtypes are more predictive of FEV1 decline and emphysema progression than original subtypes34
Peripheral eosinophil counts are linearly associated with exacerbation risk35

Figure 2 – *Summary of contributions from COPDGene to machine learning approaches to COPD subtyping. GWAS = genome-wide association study. See Figure 1 legend for expansion of other abbreviation.*

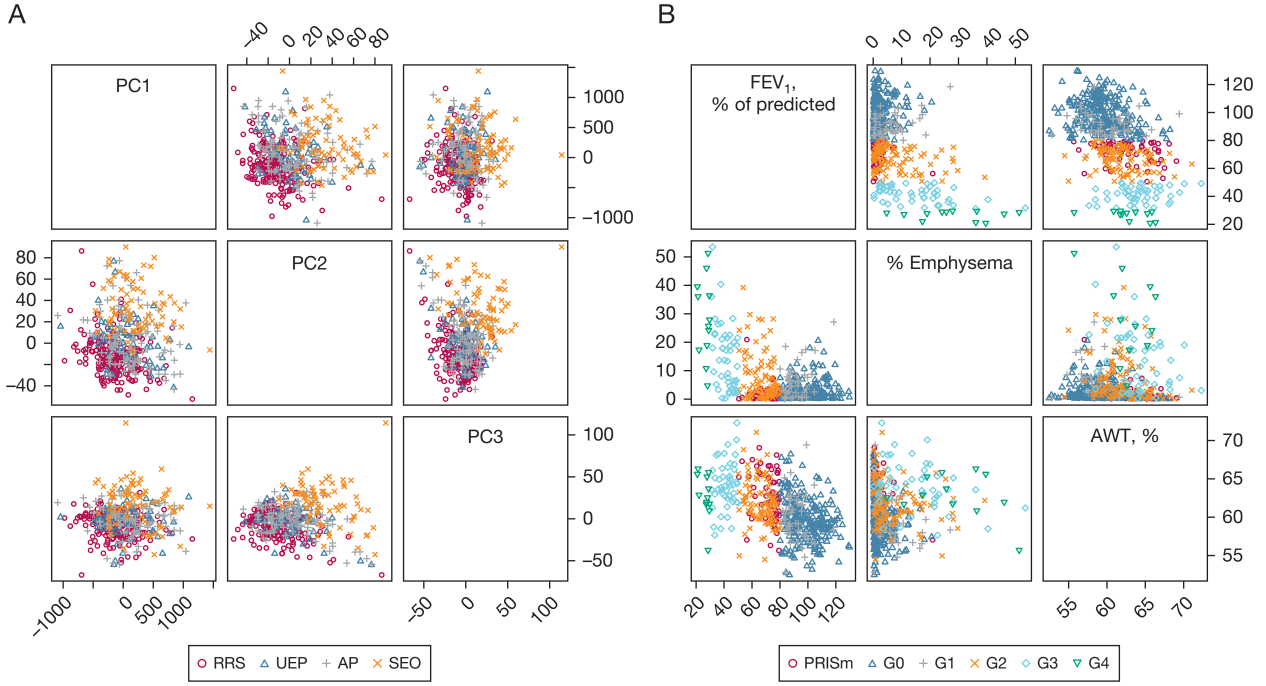

Figure 3 – *Scatterplot matrices show the distribution of clustering-defined subtypes (Castaldi et al[18]) in principal component space for 500 subjects from COPDGene (A), and the same subjects projected along the dimensions of FEV$_1$ % predicted, CT quantitative emphysema, and CT airway wall thickness with points colored by Global Initiative for Chronic Obstructive Lung Disease spirometric categories (B). AP = airway predominant; AWT, % = airway wall thickness as a percentage of total luminal area for segmental airways; G0-G4 = Global Initiative for Chronic Obstructive Lung Disease spirometric stages 0 to 4; PC = principal component; PRISm = preserved ratio impaired spirometry (ie, FEV$_1$ < 80% of predicted, FEV$_1$/FVC > 0.7); RRS = relatively resistant smokers; SEO = severe emphysema and obstruction; UEP = upper lobe emphysema predominant. See Figure 1 legend for expansion of other abbreviation.*

found two recurring clusters that seem to share characteristics with the airway-predominant COPD and severe airflow and emphysema clusters. However, Pinto et al also noted that it was not possible to perform a quantitative comparison of clustering results because the methods and variables used across studies were dissimilar, and thus quantitative assessment of the reproducibility of clustering results could not be performed.

TABLE 2] Comparison of K-Means Clustering and NBS Clustering Results Shows Little Overlap

Variable	NBS1	NBS2	NBS3	NBS4
Relatively resistant smokers	20	10	3	2
Upper lobe predominant emphysema	5	9	1	1
Airway predominant	15	11	0	3
Severe COPD	15	13	1	11

For 120 COPDGene subjects analyzed in both the Castaldi et al[18] phenotype clustering article and the Chang et al[31] gene expression clustering article, the overlap between clustering assignments was modest. Network-based stratification (NBS) clusters are ordered as in the original manuscript by average level of FEV$_1$ (ie, NBS1 has highest average FEV$_1$).

To directly address the question of the reproducibility of clustering in COPD, a collaborative study in the International COPD Genetics Consortium (ICGC) was performed to assess the subject-level similarity of clustering results from multiple methods applied across multiple cohorts.[32] This study showed that clustering results were only modestly reproducible. However, the principal component axes derived from these same datasets were very stable. This suggests that, for the set of variables studied, the COPD "phenotypic space" is a continuum rather than a group of discrete clusters. Figure 3B shows the continuous nature of the COPD phenotypic space for different sets of variables in COPDGene. This continuous phenotypic space is more amenable to dimension reduction than clustering.

A subsequent clustering reproducibility study[33] from the COPD Cohorts Collaborative International Assessment (3CIA) reported greater agreement, although the metrics of reproducibility differed between the two studies. In the ICGC study,[32] clustering reproducibility was assessed by comparing the results of clustering analyses performed de novo in each of the participating cohorts. In the 3CIA study, clustering was performed in a single cohort, and the reproducibility of cluster-specific mortality rates was assessed across multiple cohorts by using this single clustering solution. Although both studies are valid, the definition of clustering reproducibility is not the same. The ICGC results provide information on the reproducibility of the clustering process itself, whereas the 3CIA study reports the reproducibility of average characteristics and event rates of a single clustering solution.

In summary, these studies highlight that clustering is useful for identifying novel connections between clinical phenotype and molecular measures. However, the reproducibility of clustering across datasets may not be high, because COPD clinical datasets often do not have a strong clustering structure. Thus, for any clustering result, demonstration of reproducibility of the clustering process itself is essential for any claims about the generalizability or clinical translation of the cluster assignments.

What Other Machine Learning Methods Besides Clustering Can Be Used to Study COPD Heterogeneity?

As an extension of the finding that disease axes were more reproducible than clusters, Kinney et al[26] applied another dimension reduction method (factor analysis) to 28 chest CT and pulmonary function measures in COPDGene to identify COPD disease axes. In factor analysis, the contribution of the original variables to each factor can be quantified through the factor loadings for each axis. Pulmonary function measures contributed strongly to the first two factors: the first was labeled as the emphysema disease axis based on contributions from multiple CT emphysema measures, and the second was labeled as the airway disease axis due to contributions of CT measures of the thickness of the segmental airway walls. Three other factors were identified: two represented both gas trapping and hyperinflation, and one captured CT measurement variability associated with BMI. These factors were then incorporated into predictive models of mortality and clinical outcomes in COPD. Both the airway and emphysema disease axes were related to mortality, with a statistically significant, synergistic interaction between the airway and emphysema disease axes (Fig 4).

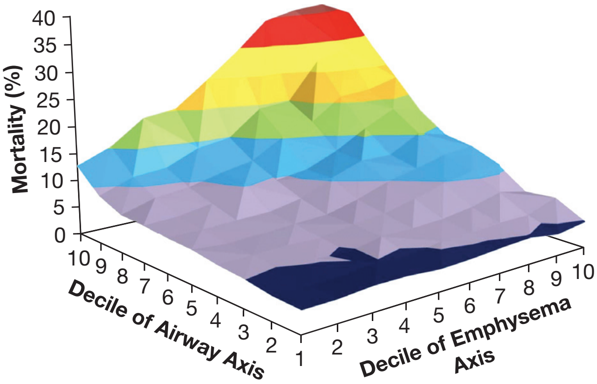

Figure 4 – *The y-axis represents the predicted probability of all-cause mortality ranging from 4% (shown in dark blue), 5% to 10% (shown in purple), 10% to 15% (shown in blue), 15% to 20% (shown in green), 20% to 25% (shown in orange), 25% to 30% (shown in yellow), 30% to 35% (shown in red), to > 35% (shown in dark red) for each decile of loading score for factors 1 (Emphysema Axis) and 2 (Airway Axis) in a Cox proportional hazards model including age, sex, current smoking, pack years of smoking, BMI, high BP, each of the five factors, the interaction between factors 1 and 2, and a quadratic term for factor 2. The x and z axes represent deciles of each axis, ranging from 1 (representing a small loading score) to 10 (representing a large loading score).*

Chen et al[34] developed an approach to generate more clinically interpretable disease axes that would allow users to have a greater level of control in determining the orientation of a disease axis. The concept of this method is to create disease axes that are oriented or “anchored” at either end by known COPD subtypes. In practice, this is done by building a logistic regression model to discriminate between the two subtypes or subgroups, with the predicted values from this model constituting a subtype-defined disease axis. We applied this method to build a chronic bronchitis disease axis. We observed that, relative to the presence or absence of chronic bronchitis at baseline, the disease axis provided better prediction for 5-year change in FEV_1 (6.4% vs 6.0% variance explained) and emphysema (12.8% vs 7.5% variance explained), and disease axis values at baseline were predictive of persistent chronic bronchitis symptoms at the COPDGene 5-year follow-up visit (Fig 5).

In summary, COPD clinical variability is typically distributed along a continuum, and continuous disease axes generated by dimension reduction methods are more natural representations of this continuum that are also more likely to be reproducible than clusters. A direct comparison between subtypes and disease axes showed that disease axes often provide more accurate prediction of future COPD-related events.

How Can Cut-Points Be Defined in a Data-Driven Way to Turn Continuous COPD Measures Into Subtypes?

If continuous disease axes are more accurate and reproducible than clusters, how could such continuous phenotypes be used to help make clinical decisions?

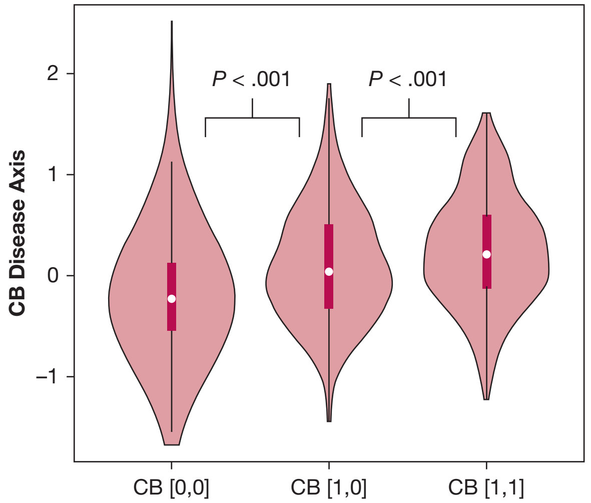

Figure 5 – *Distribution of chronic bronchitis disease axis values at the COPDGene baseline visit according to presence of chronic bronchitis symptoms at the baseline and 5-year study visit. Subjects with persistent chronic bronchitis symptoms (ie, present at both visits, CB [1,1]) had disease axis values that were higher than subjects without chronic bronchitis (CB [0,0]) and subjects with intermittent symptoms (CB [1,0] for chronic bronchitis at baseline but not at the 5-year visit).* P *values were calculated by using the Mann-Whitney* U *test. See Figure 1 legend for expansion of abbreviation.*

Yun et al[35] addressed this question by examining the relation between peripheral blood eosinophil measurements and risk of COPD exacerbations. In COPDGene subjects in GOLD spirometric stages 2, 3, or 4, the number of respiratory exacerbations was linearly related to the number of eosinophils in the peripheral blood, and this relation was stronger with absolute eosinophil counts rather than with eosinophil percentage. To determine a reasonable cutoff, prediction models for exacerbations were made using a range of cutoffs on absolute eosinophil count, with a value of 300 cells/μL having the best performance. These models were validated in subjects from the ECLIPSE study. These findings are consistent with other reports, including an analysis of 7,225 subjects with COPD in the Copenhagen General Population Study, which also found that absolute eosinophil counts provided superior prediction of respiratory exacerbations relative to eosinophil percentages.[36] This study used a similar count threshold of 340 cells/μL. Another analysis of 7,245 subjects with COPD confirmed that a cutoff of 300 cells/μL was associated with exacerbation rate in multivariate models, and the exacerbation rate increased with higher cutoff thresholds.[37]

The study by Yun et al[35] provides a roadmap for how to turn continuous COPD phenotypes (in this case, peripheral eosinophilia) into clinically relevant subtypes according to criteria based on assessment of risk for COPD-related outcomes. This article shows how predictive models can be used to identify specific subtype cutoffs, although this method also raises the possibility of having different sets of COPD subtypes corresponding to different clinical outcomes.

How Can Machine Learning on Chest CT Data Improve Our Ability to Characterize COPD Heterogeneity?

Semi-automated classification of emphysema patterns and airway wall thickness from thousands of COPDGene CT scans has improved our ability to divide COPD into distinct subgroups. Mendoza et al[38] used k-nearest neighbor clustering to quantify distinct CT emphysema patterns by comparing local lung density histograms vs a set of manually curated reference patterns of pathologic emphysema in > 9,000 CT scans from COPDGene. The resulting local histogram emphysema quantifications had stronger associations to a range of spirometric and functional measures than standard measures of CT emphysema,[39] and genome-wide association study of these measures identified known and novel genetic associations.[40] One of the genetic regions identified by the genome-wide association study was subsequently shown using CRISPR gene editing to contain a fibroblast-specific enhancer element that increases the expression of *TGFB2* in fibroblasts; this finding provides additional genetic evidence of the link between emphysema and TGF-β signaling in human COPD.[41]

Are There Distinct Trajectories of Lung Function Over the Life Course That Correspond to Molecular Subtypes of COPD?

COPD subtypes are usually defined based on cross-sectional data, but subtypes learned in this manner can be confounded by differences in disease severity. For certain tasks, such as the identification of genetic associations to COPD, it is desirable to identify distinct patterns of disease progression that are not confounded by these severity differences. To address this need, Ross et al[42] developed a Bayesian modeling approach that incorporates the concept of disease trajectories into COPD subtype identification. This study used decades-long longitudinal spirometric data in the Normative Aging Study (NAS) to identify and model four distinct patterns of FEV_1 decline. Interestingly, the trajectory with the most rapid rate of decline in mid-life was also characterized by the lowest maximal FEV_1 attained, suggesting this was a low lung growth/rapid decline trajectory (Fig 6).

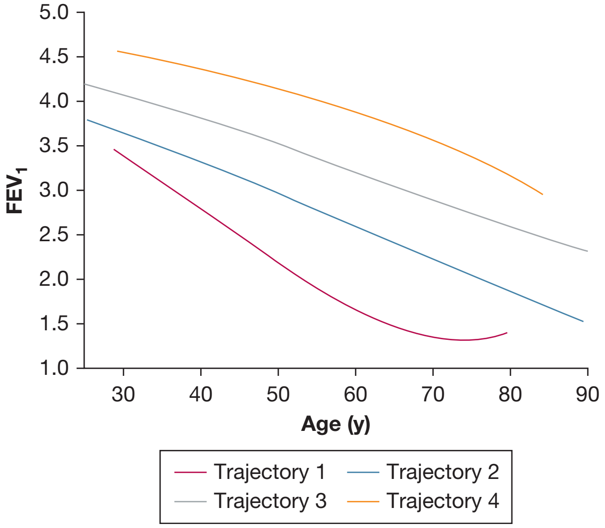

Figure 6 – *Four lung function trajectories learned from analyzing 1,060 men followed up for > 20 years in the Normative Aging Study. Trajectory 1 was characterized by both a lower maximal FEV_1 attained as well as a more rapid rate of lung function loss in mid-life. The other trajectories differed primarily in maximal FEV_1 attained but not in rate of decline. See Figure 1 legend for expansion of abbreviation.*

These models were then applied to a subset of COPDGene subjects to infer their lung function trajectory assignment. In COPDGene, subjects with severe COPD were overrepresented in the low growth/rapid decline trajectory. This trajectory seems to be strongly associated with genetic differences based on a higher rate of parental COPD and the high genetic contribution to trajectories identified from heritability-based analysis.[43] These findings are consistent with the results of other trajectory-based analyses of COPD,[44-46] and this is a promising approach for integrating information between studies that have varying amounts of longitudinal follow-up available.

Discussion and Future Directions

The main findings from the studies covered in this review are as follows: (1) clustering is most useful for exploratory analyses of COPD subtypes; (2) continuous disease axes more accurately represent COPD heterogeneity than clusters; (3) chest CT phenotypes obtained through machine learning algorithms have improved our ability to quantify COPD heterogeneity and have led to novel biological discoveries, including in the TGF-β pathway; and (4) trajectories of lung growth and decline show strong genetic influences and may enable more powerful biological discoveries in COPD.

Although the use of machine learning with rich COPD datasets is promising, a strict replication analysis in 10 cohorts found that clustering results were poorly reproducible. The conclusion from this study is that, in some instances, clustering is poorly suited for COPD data that are distributed along a continuum without distinct subgroups.[32] Because of these issues of reproducibility, greater focus has been placed on the identification of continuous measures of COPD-related disease processes, such as treatable traits and disease axes. Disease axes have been shown to be more reproducible than clusters[32] and more predictive of 5-year changes in FEV_1 and emphysema.[34]

Clinical translation of disease axes and treatable traits requires that clinically relevant cutoffs be identified for these continuous measures. The research by Yun et al[35] in peripheral eosinophilia shows how support for cutoff values can be derived from predictive risk models. By relating eosinophilia to exacerbation risk, standard statistical methods provided support for a cutoff of 300 cells/μL. Based on many additional studies of stability of blood eosinophil counts and retrospective analysis of clinical trial data, the GOLD 2019 criteria also included the 300 eosinophils/μL threshold for considering first-line inhaled corticosteroids in subjects with group D COPD.[2] Thus, peripheral eosinophilia is a concrete example of how a continuous COPD phenotype can be translated into subtypes for clinical practice through the development and replication of risk models for a COPD-related outcome. This implies that different cutoffs and subgroups may need to be defined for different outcomes. Thus, rather than asking "What are the subtypes of COPD?" it may be better to determine which subtypes are the most useful for a specific clinical purpose.

As we discover more COPD-related biomarkers, we can expect that COPD subtypes will increasingly be defined by using a combination of clinical features, imaging characteristics, and molecular markers. As our knowledge of genetic associations to COPD steadily increases,[47] and the quality of COPD phenotypes improves, updated COPD subtype definitions will better capture the clinical and biological heterogeneity of COPD.

What are the key areas in which we anticipate additional contributions from COPDGene? First, when 10-year follow-up data are available, associations to disease progression will be more apparent, and more detailed descriptions of lung function trajectories will be possible. Second, the large-scale generation of DNA sequencing, RNA sequencing, DNA methylation, and proteomic data from blood samples at the 5- and 10-year visits will identify key molecular biomarkers of COPD progression that will lead to improved definitions of COPD molecular subtypes. Third, updated analyses of disease progression can identify the minimal sets of variables necessary for accurate risk stratification, making subtyping more broadly applicable in a clinical setting. Finally, advances in machine learning methods may lead to a more detailed understanding of the relation between COPD heterogeneity and disease progression.

Acknowledgements

Financial/nonfinancial disclosures: The authors have reported to *CHEST* the following: P. J. C. has received research support and consulting fees from GlaxoSmithKline and Novartis. E. K. S. received honoraria from Novartis for Continuing Medical Education Seminars; and grant and travel support from GlaxoSmithKline. C. P. H. reports personal fees from Mylan, AstraZeneca, Concert Pharmaceuticals, and 23andMe; and grants from Novartis and Boehringer Ingelheim. G. W. reports grants and other support from Boehringer Ingelheim, PulmonX, BTG Interventional Medicine, Janssen Pharmaceuticals, and GlaxoSmithKline. R. S. J. E. reports personal fees from Boehringer Ingelheim, Eolo Medical, and Toshiba. M. H. C. reports grants from GlaxoSmithKline; and personal fees from Genentech. None declared (A. B., J. Y., J. C. R., G. L. K., K. A. Y., E. A. R., D. A. L., G. J. C., J. G. D., S. I. R., R. C., B. J. M., J. C., J. E. H.).

***COPDGene Investigators, Core Units:** *Administrative Center*: James D. Crapo, MD (Principal Investigator); Edwin K. Silverman, MD, PhD (Principal Investigator); Barry J. Make, MD; and Elizabeth A. Regan, MD, PhD. *Genetic Analysis Center*: Terri Beaty, PhD; Ferdouse Begum, PhD; Peter J. Castaldi, MD; Michael Cho, MD; Dawn L. DeMeo, MD, MPH; Adel R. Boueiz, MD; Marilyn G. Foreman, MD, MS; Eitan Halper-Stromberg, MD, PhD; Lystra P. Hayden, MD; Craig P. Hersh, MD, MPH; Jacqueline Hetmanski, MS, MPH; Brian D. Hobbs, MD; John E. Hokanson, MPH, PhD; Nan Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD; Margaret M. Parker, PhD; Dmitry Prokopenko, PhD; Dandi Qiao, PhD; Elizabeth A. Regan, MD, PhD; Phuwanat Sakornsakolpat, MD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD; and Sungho Won, PhD. *Imaging Center*: Juan Pablo Centeno, MSc; Jean-Paul Charbonnier, PhD; Harvey O. Coxson, PhD; Craig J. Galban, PhD; MeiLan K. Han, MD; Eric A. Hoffman, PhD; Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber, BA; David A. Lynch, MB; Pietro Nardelli, PhD; John D. Newell Jr, MD; Aleena Notary, MS; Andrea Oh, MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren, MHA; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; Gonzalo Vegas Sanchez-Ferrero, PhD; Lucas Veitel, BA; George R. Washko, MD; and Carla G. Wilson, MS. *PFT QA Center, Salt Lake City, UT*: Robert Jensen, PhD. *Data Coordinating Center and Biostatistics*, *National Jewish Health, Denver, CO*: Douglas Everett, PhD; Jim Crooks, PhD; Katherine Pratte, PhD; Matt Strand, PhD; and Carla G. Wilson, MS. *Epidemiology Core*, *University of Colorado Anschutz Medical Campus, Aurora, CO*: John E. Hokanson, MPH, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; and Kendra A. Young, PhD. *Mortality Adjudication Core:* Surya P. Bhatt, MD; Jessica Bon, MD; Alejandro A. Diaz, MD, MPH; MeiLan K. Han, MD; Barry Make, MD; Susan Murray, ScD; Elizabeth Regan, MD; Xavier Soler, MD; and Carla G. Wilson, MS. *Biomarker Core*: Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; and Farnoush Banaei-Kashani, PhD.

Other contributions: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

References

1. Rennard SI, Vestbo J. The many "small COPDs." *Chest.* 2008;134(3):623.
2. Singh D, Agustí AGN, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease: the GOLD science committee report 2019. *Eur Respir J.* 2019;53(5):1900164.
3. Bhatt SP, Washko GR, Hoffman EA, et al. Imaging advances in chronic obstructive pulmonary disease. Insights from the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease (COPDGene) Study. *Am J Respir Crit Care Med.* 2019;199(3):286-301.
4. Stringer WW, Porsasz J, Bhatt SP, McCormack MC, Make BJ, Casaburi R. Physiologic insights from the COPDGene study. *Journal of the COPD Foundation.* 2019;6(3):256-266.
5. Maselli DJ, Bhatt SP, Anzueto A, et al. Clinical epidemiology of COPD: insights from 10 years of the COPDGene study. *Chest.* 2019;156(2):228-238.
6. Ragland MF, Benway CJ, Lutz SM, et al. Genetic advances in chronic obstructive pulmonary disease. Insights from COPDGene. *Am J Respir Crit Care Med.* 2019;200(6):677-690.
7. Regan EA, Hersh CP, Castaldi PJ, et al. Omics and the search for blood biomarkers in chronic obstructive pulmonary disease. Insights from COPDGene. *Am J Respir Cell Mol Biol.* 2019;61(2):143-149.
8. Fletcher CM, Gilson JG, Hugh-Jones P, Scadding JG. Terminology, definitions, and classification of chronic pulmonary emphysema and related conditions: a report of the conclusions of a Ciba Guest Symposium. *Thorax.* 1959;14(4):286-299.
9. Burrows B, Niden AH, Fletcher CM, Jones NL. Clinical types of chronic obstructive lung disease in London and in Chicago. A study of one hundred patients. *Am Rev Respir Dis.* 1964;90:14-27.
10. Burrows B, Fletcher CM, Heard BE, Jones NL, Wootliff JS. The emphysematous and bronchial types of chronic airways obstruction. A clinicopathological study of patients in London and Chicago. *Lancet.* 1966;1(7442):830-835.
11. Donaldson GC, Seemungal TAR, Bhowmik A, Wedzicha JA. Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease. *Thorax.* 2002;57(10):847-852.
12. Hurst JR, Vestbo J, Anzueto A, et al. Susceptibility to exacerbation in chronic obstructive pulmonary disease. *N Engl J Med.* 2010;363(12):1128-1138.
13. Gibson PG, Simpson JL. The overlap syndrome of asthma and COPD: what are its features and how important is it? *Thorax.* 2009;64(8):728-735.
14. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med.* 2003;348(21):2059-2073.
15. Pistolesi M, Camiciottoli G, Paoletti M, et al. Identification of a predominant COPD phenotype in clinical practice. *Respir Med.* 2008;102(3):367-376.
16. Garcia-Aymerich J, Gómez FP, Benet M, et al. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax.* 2011;66(5):430-437.
17. Cho M, Washko GR, Hoffmann TJ, et al. Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation. *Respir Res.* 2010;11:30.
18. Castaldi PJ, Dy JG, Ross J, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax.* 2014;69(5):415-422.
19. Vanfleteren L, Spruit M, Groenen M, et al. Clusters of comorbidities based on validated objective measurements and systemic inflammation in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med.* 2013;187(7):728-735.

20. Burgel PR, Paillasseur JL, Caillaud D, et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J*. 2010;36(3):531-539.

21. Pinto LM, Alghamdi M, Benedetti A, Zaihra T, Landry T, Bourbeau J. Derivation and validation of clinical phenotypes for COPD: a systematic review. *Respir Res*. 2015;16(1):50.

22. Woodruff PG, Agustí AGN, Roche N, Singh D, Martinez FJ. Current concepts in targeting chronic obstructive pulmonary disease pharmacotherapy: making progress towards personalised management. *Lancet*. 2015;385(9979):1789-1798.

23. Woodruff PG, Modrek B, Choy DF, et al. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am J Respir Crit Care Med*. 2009;180(5):388-395.

24. Agustí AGN, Bel E, Thomas M, et al. Treatable traits: toward precision medicine of chronic airway diseases. *Eur Respir J*. 2016;47(2):410-419.

25. McCarthy MI. Painting a new picture of personalised medicine for diabetes. *Diabetologia*. 2017;60(5):793-799.

26. Kinney GL, Santorico SA, Young KA, et al. Identification of chronic obstructive pulmonary disease axes that predict all-cause mortality: the COPDGene study. *Am J Epidemiol*. 2018;187(10):2109-2116.

27. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2010;7(1):32-43.

28. Boueiz A, Lutz SM, Cho M, et al. Genome-wide association study of the genetic determinants of emphysema distribution. *Am J Respir Crit Care Med*. 2017;195(6):757-771.

29. Boueiz A, Pham B, Chase R, et al. Integrative genomics analysis identifies ACVR1B as a candidate causal gene of emphysema distribution. *Am J Respir Cell Mol Biol*. 2019;60(4):388-398.

30. Boueiz A, Chang Y, Cho M, et al. Lobar Emphysema distribution is associated with 5-year radiological disease progression. *Chest*. 2017;153(1):65-76.

31. Chang Y, Glass K, Liu YY, et al. COPD subtypes identified by network-based clustering of blood gene expression. *Genomics*. 2016;107(2-3):51-58.

32. Castaldi PJ, Benet M, Petersen H, et al. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax*. 2017;72(11):998-1006.

33. Burgel PR, Paillasseur JL, Janssens W, et al. A simple algorithm for the identification of clinical COPD phenotypes. *Eur Respir J*. 2017;50(5):1701034.

34. Chen J, Cho M, Silverman EK, et al. Turning subtypes into disease axes to improve prediction of COPD progression. *Thorax*. 2019;74(9):906-909.

35. Yun JH, Lamb A, Chase R, et al. Blood eosinophil count thresholds and exacerbations in patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol*. 2018;141(6):2037-2047.e10.

36. Vedel-Krogh S, Nielsen SF, Lange P, Vestbo J, Nordestgaard BG. Blood eosinophils and exacerbations in chronic obstructive pulmonary disease. The Copenhagen General Population Study. *Am J Respir Crit Care Med*. 2016;193(9):965-974.

37. Zeiger RS, Tran TN, Butler RK, et al. Relationship of blood eosinophil count to exacerbations in chronic obstructive pulmonary disease. *J Allergy Clin Immunol Pract*. 2018;6(3):944-954.e5.

38. Mendoza CS, Washko GR, Crapo JD, et al. Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions. *Proc IEEE Int Symp Biomed Imaging*. 2012:474-477.

39. Castaldi PJ, San José Estépar R, Mendoza CS, et al. Distinct quantitative CT emphysema patterns are associated with physiology and function in smokers. *Am J Respir Crit Care Med*. 2013;188(9): 1083-1090.

40. Castaldi PJ, Cho M, San José Estépar R, et al. Genome-wide association identifies regulatory loci associated with distinct local histogram emphysema patterns. *Am J Respir Crit Care Med*. 2014;190(4):399-409.

41. Parker MM, Hao Y, Guo F, et al. Identification of an emphysema-associated genetic variant near TGFB2 with regulatory effects in lung fibroblasts. *Elife*. 2019;8.

42. Ross JC, Castaldi PJ, Cho M, et al. A Bayesian nonparametric model for disease subtyping: application to emphysema phenotypes. *IEEE Trans Med Imaging*. 2017;36(1):343-354.

43. Ross JC, Castaldi PJ, Cho M, et al. Longitudinal modeling of lung function trajectories in smokers with and without chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2018;198(8):1033-1042.

44. Lange P, Celli B, Agustí AGN, et al. Lung-function trajectories leading to chronic obstructive pulmonary disease. *N Engl J Med*. 2015;373(2):111-122.

45. Bui DS, Lodge CJ, Burgess JA, et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. *Lancet Respir Med*. 2018;6(7):535-544.

46. Agustí AGN, Faner R. Lung function trajectories in health and disease. *Lancet Respir Med*. 2019;7(4):358-364.

47. Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nature Genetics*. 2019;51(3):494-505.