Polygenic and transcriptional risk scores identify chronic obstructive pulmonary disease subtypes in the COPDGene and ECLIPSE cohort studies

Matthew Moll,[a,b,c,x] Julian Hecker,[a,x] John Platig,[d] Jingzhou Zhang,[e] Auyon J. Ghosh,[f] Katherine A. Pratte,[g] Rui-Sheng Wang,[a] Davin Hill,[h] Iain R. Konigsberg,[i] Joe W. Chiles, III[j] Craig P. Hersh,[a,b,x] Peter J. Castaldi,[a,j,k,x] Kimberly Glass,[a,x] Jennifer G. Dy,[h] Don D. Sin,[l] Ruth Tal-Singer,[m] Majd Mouded,[n] Stephen I. Rennard,[o] Gary P. Anderson,[p] Gregory L. Kinney,[q] Russell P. Bowler,[r] Jeffrey L. Curtis,[s,t] Merry-Lynn McDonald,[i,u,v] Edwin K. Silverman,[a,b,x] Brian D. Hobbs,[w,y] and Michael H. Cho[a,b,x,y,*]

[a]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, 02115, USA
[b]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, 02115, USA
[c]Division of Pulmonary, Critical Care, Sleep and Allergy, Veterans Affairs Boston Healthcare System, West Roxbury, MA, 02123, USA
[d]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, 22903, USA
[e]The Pulmonary Center, Boston University Medical Center, Boston, MA 02118, USA
[f]Division of Pulmonary, Critical Care, and Sleep Medicine, SUNY Upstate Medical University, Syracuse, NY, 13210, USA
[g]Department of Biostatistics, National Jewish Health, Denver, CO, 80206, USA
[h]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, 02115, USA
[i]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, 80045, USA
[j]Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, 35233, USA
[k]Division of General Internal Medicine and Primary Care, Department of Medicine, Brigham and Women's Hospital, Boston, MA, 02115, USA
[l]Centre for Heart Lung Innovation, St. Paul's Hospital, and Department of Medicine (Respiratory Division), University of British Columbia, Vancouver, BC, Canada
[m]Global Allergy and Airways Patient Platform, Vienna, Austria
[n]Novartis Institute for Biomedical Research, Cambridge, MA, USA
[o]Division of Pulmonary, Critical Care, and Sleep Medicine, University of Nebraska, Omaha, NE, 68198, USA
[p]Lung Health Research Centre, Department of Biochemistry and Pharmacology, University of Melbourne, Melbourne, Victoria, Australia
[q]Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, 80045, USA
[r]Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Health, Denver, CO, 80206, USA
[s]Division of Pulmonary and Critical Care Medicine, University of Michigan School of Medicine, Ann Arbor, MI, 48109, USA
[t]Medical Service, Veterans Affairs Ann Arbor Healthcare System, Ann Arbor, MI, 48109, USA
[u]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, 701, 19th Street S., LHRB 440, Birmingham, AL, 35233, USA
[v]Department of Genetics, University of Alabama at Birmingham, Birmingham, AL, 35233, USA
[w]Regeneron Pharmaceutical, Tarrytown, NY, USA
[x]Harvard Medical School, Boston, MA, 02115, USA

eBioMedicine
2024;110: 105429
Published Online xxx
https://doi.org/10.1016/j.ebiom.2024.105429

Summary

Background Genetic variants and gene expression predict risk of chronic obstructive pulmonary disease (COPD), but their effect on COPD heterogeneity is unclear. We aimed to define high-risk COPD subtypes using genetics (polygenic risk score, PRS) and blood gene expression (transcriptional risk score, TRS) and assess differences in clinical and molecular characteristics.

Methods We defined high-risk groups based on PRS and TRS quantiles by maximising differences in protein biomarkers in a COPDGene training set and identified these groups in COPDGene and ECLIPSE test sets. We tested multivariable associations of subgroups with clinical outcomes and compared protein–protein interaction networks and drug repurposing analyses between high-risk groups.

Findings We examined two high-risk omics-defined groups in non-overlapping test sets (n = 1133 NHW COPDGene, n = 299 African American (AA) COPDGene, n = 468 ECLIPSE). We defined "high activity" (low PRS, high TRS) and

*Corresponding author. 181 Longwood Avenue, Boston, MA, 02115, USA.
E-mail address: remhc@channing.harvard.edu (M.H. Cho).
[y]Jointly supervised and co-senior authors.

"severe risk" (high PRS, high TRS) subgroups. Participants in both subgroups had lower body-mass index (BMI), lower lung function, and alterations in metabolic, growth, and immune signalling processes compared to a low-risk (low PRS, low TRS) subgroup. "High activity" but not "severe risk" participants had greater prospective FEV_1 decline (COPDGene: −51 mL/year; ECLIPSE: −40 mL/year) and proteomic profiles were enriched in gene sets perturbed by treatment with 5-lipoxygenase inhibitors and angiotensin-converting enzyme (ACE) inhibitors.

Interpretation Concomitant use of polygenic and transcriptional risk scores identified clinical and molecular heterogeneity amongst high-risk individuals. Proteomic and drug repurposing analysis identified subtype-specific enrichment for therapies and suggest prior drug repurposing failures may be explained by patient selection.

Funding National Institutes of Health.

Keywords: Polygenic risk scores; COPD; Transcriptomics; Endotyping; Drug repurposing

Research in context

Evidence before this study

Genetic variants and gene expression have been previously associated with the risk of developing chronic obstructive pulmonary disease (COPD). However, their role in defining the heterogeneity of COPD subtypes has not been fully explored.

Added value of this study

We utilised both polygenic (PRS) and transcriptional (TRS) risk scores to identify high-risk COPD subtypes. This approach highlighted two subgroups: "high activity" and "severe risk." The findings demonstrate distinct clinical and molecular characteristics in these subgroups, with differences in body-mass index (BMI), lung function, and proteomic profiles. Moreover, proteomic and drug repurposing analyses revealed subtype-specific enrichment for certain therapies, such as 5-lipoxygenase inhibitors and ACE inhibitors, which could explain previous failures in drug repurposing due to patient selection issues.

Implications of all the available evidence

Integrating polygenic and transcriptional risk scores provides a more nuanced understanding of COPD heterogeneity. These findings suggest that patient stratification using omics-based approaches could enhance the effectiveness of targeted therapies, potentially leading to better clinical outcomes for high-risk COPD subtypes.

Introduction

Chronic obstructive pulmonary disease (COPD) is a leading cause of morbidity and mortality worldwide.[1] Although COPD is characterised by irreversible airflow obstruction, there is marked heterogeneity amongst individuals in emphysema and airway pathology, exacerbation incidence, and lung function decline.[2,3] Identifying individuals at high risk for rapid COPD progression or eventual severe disease is critically important to implement personalised therapeutic approaches.

Large-scale omics data offer the potential to identify, via a simple blood test, high risk groups that share distinct, targetable pathobiology. Genetics, quantified with polygenic risk scores (PRSs), can identify individuals at high risk for coronary artery disease and guide consideration of statin therapy earlier than advised by current guidelines.[4] In cancer, integrating genetic and transcriptomic profiling can improve therapy recommendations and outcomes.[5] We demonstrated that both a PRS and a transcriptional risk score (TRS) independently predict COPD.[6,7] The TRS predicted COPD with an odds ratio of 3.3 and area-under-the-curve of 0.79.[7] Although the PRS and TRS were both based on spirometry measures, the scores are not correlated[7] and likely capture different aspects of lung pathobiology. Specifically, COPD genetic risk loci are enriched for aspects of lung development and have a greater effect in early COPD[6,8]; in contrast, the COPD TRS is associated with markers of inflammation and lung function decline and may reflect disease activity and propensity toward disease progression.[7] Thus, it may be possible to leverage the different features of the PRS and TRS to identify clinically and biologically distinct COPD subtypes.

Despite advances in omics-based risk prediction, important clinical translation questions remain. Omics risk scores are usually standardised for statistical analyses, leaving the question of how to use them to risk-stratify individuals.[9] Risk scores are also continuous measures, often normally distributed; the issue of attempting to identify subtypes along a continuum has been previously recognised.[10] Despite this limitation, there is a need to classify individuals to link omics-defined high-risk groups, which might benefit from specific therapies, with specific pathobiological

processes and treatment decisions. COPD drug and drug repurposing candidates have high failure rates in clinical trials,[11] but it remains unknown if these therapies have failed because of patient selection, which currently does not utilise omics or other biomarkers.

The PRS was effective for predicting COPD severity and incident COPD, while the TRS was better at predicting FEV_1 decline; combining both risk scores may identify subgroups at risk for multiple important COPD outcomes. Therefore, we hypothesised that our published PRS and TRS, both based on spirometry, could identify COPD subtypes (i.e., heterogeneity) within high-risk groups with clinical and biological differences in two cohorts of ever-smokers. We aimed to develop an approach for how omics risk scores can be applied to populations and leveraged for precision medicine. We used proteomics to obtain an additional biological view of omics-defined subgroups and performed in silico drug repurposing analyses to identify potential subgroup-specific drug repurposing candidates.

Methods

Study populations

COPDGene

We included Genetic Epidemiology of COPD (COPDGene) study (ClinicalTrials.gov Identifier: NCT00608764) participants with single nucleotide polymorphism (SNP) genotyping, RNA-sequencing, and SomaScan proteomic data to calculate the PRS, TRS, and performed differential protein expression analyses, respectively. Briefly, the COPDGene study recruited n = 10,198 non-Hispanic white (NHW) and African American (AA) individuals aged 45–80 years with ≥10 pack-years of smoking history.[12] COPDGene began as a cross-sectional case–control study that was extended into a longitudinal study including 5 and 10-year follow up visits. Anthropometric, spirometry, and computed tomography (CT) imaging measures were performed at each visit. We obtained genotype data at baseline, and RNA-sequencing and SomaScan proteomic data at the 5-year follow up visit.

Single nucleotide polymorphism (SNP) genotyping was performed using the Illumina (San Diego, CA) HumanOmniExpress array. Genotyping at the Z and S alleles was performed and participants with severe alpha-1 antitrypsin deficiency were excluded. Imputation was performed using the Michigan Imputation Server to the Haplotype Reference Consortium[13] and 1000 Genomes Phase I v3 Cosmopolitan reference panels, for non-Hispanic whites and African Americans, respectively. Variants with an r^2 value of ≤0.3 were removed.

ECLIPSE

As previously described,[7] we included Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE) study (ClinicalTrials.gov Identifier: NCT00292552) participants with SNP genotyping data, whole blood microarray data, and at least two FEV_1 measurements. The ECLIPSE study recruited n = 2140 individuals with COPD aged 40–75 years and ≥10 pack-years of smoking history.[14] Baseline anthropometric, spirometry, and CT imaging measures were collected. Blood samples were also collected at study enrolment and, for a subset of samples, both genotype and gene expression microarray data are available. If individuals met the spirometry criteria for Global Initiative for Chronic Obstructive Lung Disease (GOLD)[15] stage 2–4 COPD at enrolment, they returned every six months for three years for repeat spirometry.

In ECLIPSE, SNP genotyping was performed using the Illumina HumanHap 550 V3 (Illumina, San Diego, CA) array. Subjects and markers with call rates <95% were excluded. Imputation was performed using the Michigan Imputation Server and Haplotype Reference Consortium[13] reference panel.

Cohort expression data

COPDGene: RNA sequencing data. At the 5-year follow up of the COPDGene study, whole blood was obtained and stored in PAXgene Blood RNA tubes. Collection and processing of RNA-sequencing data was previously described.[7,16] Briefly, total RNA was extracted with the Qiagen PreAnalytiX PAXgene Blood miRNA Kit (Qiagen, Valencia, CA). After undergoing quality assurance, samples were globin reduced and cDNA library preparation was performed; 75 bp reads with a mean of 20 million reads per sample were generated using an Illumina HiSeq 2500. Count data were filtered to include transcripts with >1 count per million (CPM) in 99% of samples, and were subsequently normalised by log-CPM transformation using the edgeR R package.[17] Counts were adjusted for library depth, and batch effects were removed using the limma removeBatchEffects function.[18]

ECLIPSE: microarray data. ECLIPSE participants had blood samples collected at the time of study enrolment, and total RNA was extracted using PAXgene Blood miRNA kits and hybridised to the Affymetrix Human Gene 1.1 ST array. If transcripts were represented by multiple probes, we chose the probe with the greatest interquartile range. Batch effects were removed using the limma removeBatchEffects function.[18] Our prior publication[7] contains further details regarding preparation and processing of RNA data.

Prior to analyses, we limited transcripts to those present in both data sets based on HGNC symbols. For ECLIPSE microarray data, some gene transcripts were represented by multiple probes. In these cases, we chose the probe with the greatest interquartile range. We also scaled the RNAseq count and microarray gene expression data to have a mean of 0 and standard deviation of 1.

Proteomic data: COPDGene

Blood proteomic data were measured using SomaScan v4.0, which uses aptamers (i.e., SOMAmers) to quantify 4776 unique human proteins. Using the SomaScan 5K platform, we performed plate hybridisation, median signal normalisation, and plate scaling and calibration of SOMAmers to control for variability across array signals, inter-run variability, inter-assay variation between analytes and batch differences between plates. Further details regarding SomaScan data and preparation have been previously published.[19]

Polygenic and transcriptional risk scores

The COPD PRS and TRS were both based on spirometry and previously described.[6,7]

Polygenic risk score

We previously published a PRS[6] using GWASs of FEV_1 and FEV_1/FVC performed in approximately 500,000 individuals from the UK Biobank and SpiroMeta consortium.[20] We calculated PRSs for FEV_1 and FEV_1/FVC separately using lassosum,[21] a penalised regression approach that minimises collinearity, provides feature selection, and accounts for linkage disequilibrium. We summed the FEV_1 and FEV_1/FVC scores into a composite risk score, as previously performed.[6] COPDGene and ECLIPSE were not used to develop the PRS and represent external datasets.

To maximise our ability to separate subjects based on the genetic risk score and to minimise potential confounding by genetic ancestry, we analysed NHW and AA participants separately in this analysis and residualised by regressing out principal components of genetic ancestry from the PRS before use.

Transcriptional risk score

We previously published a TRS[7] in a training sample of COPDGene using least absolute shrinkage and selection operator (LASSO) penalised regression[22] in 1374 individuals from the COPDGene study and tested its performance in a held-out sample of 674 individuals. We have since obtained RNA-sequencing data (TOPMed Freeze 4) in an additional 459 NHW and 143 AA participants and have added these participants to the COPDGene testing set. We ensured that none of the samples used in the training of the TRS were included in our COPDGene testing set.

Statistics

Overview of study design

To identify high-risk subgroups based on continuous scores, we used the same previously defined COPDGene training set,[7] and tested among PRS and TRS quantiles to maximise the number of associated differentially expressed proteins across the resulting subtype partitions (Fig. 1). We included only Europeans in the

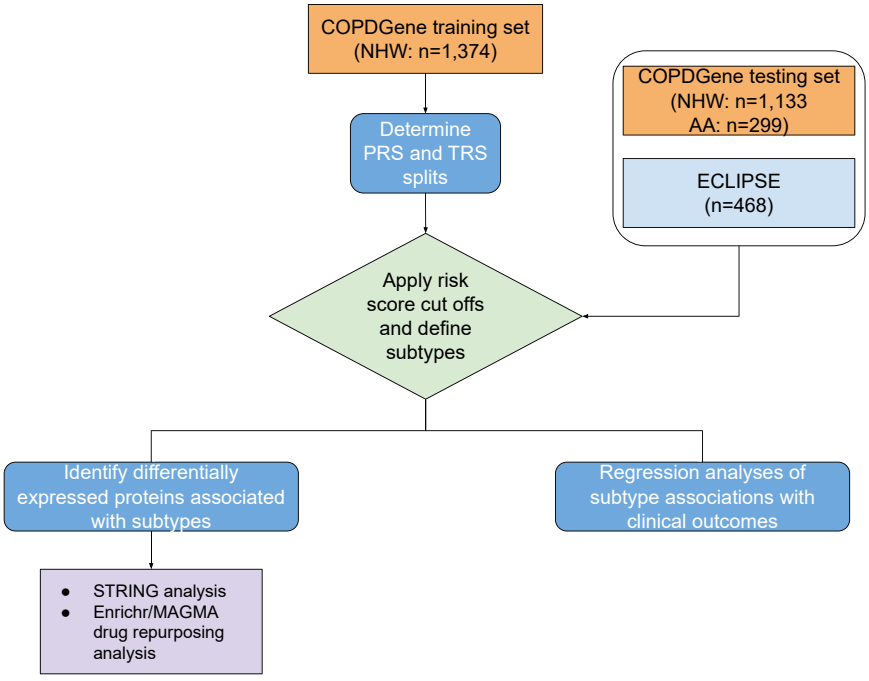

Fig. 1: Schematic of study design. COPD, chronic obstructive pulmonary disease. COPDGene, Genetic Epidemiology of COPD study. ECLIPSE, Evaluation of COPD to Longitudinally Identify Predictive Surrogate Endpoints study. PRS, polygenic risk score. TRS, transcriptional risk score. STRING, Search Tool for the Retrieval of Interacting Genes/Proteins. MAGMA, Multi-marker Analysis of GenoMic Annotation.

training dataset as the PRS was developed in European ancestry individuals. We then determined the raw (non-standardised) score cut offs associated with the corresponding percentile values in the COPDGene training set. This approach facilitated classifying each participant in the COPDGene testing set and the external ECLIPSE validation set into an omics-defined subtype. We characterised the newly defined subtypes using proteomic network and drug repurposing analyses, and applied multivariable linear regressions to test the association of subtypes with COPD-related outcomes.

Determining risk score divisions and identifying omics-defined subtypes

Omics-based risk scores are typically standardised prior to statistical analysis, and therefore, have a normal distribution and by design do not lend themselves to clustering analyses (Figure S1). Yet, for clinical application, patients need to be categorised into groups. First, to determine whether one or more clusters are optimal, we calculated the gap statistic based on the PRS and TRS using the clusGap function (cluster R package) with a maximum of 8 k means clusters and 500 bootstrap iterations. The gap statistic provides a measure of dispersion for each cluster and compares this dispersion metric to the expected dispersion under the null distribution[23]; thus, the difference (or "gap") between the observed and expected within cluster dispersion is used to calculate the gap statistic and the maximum gap statistic over a range of cluster numbers indicates the optimal number of clusters.

As an alternative to clustering, individuals are commonly placed into omics risk-score quantiles, and participants in each quantile are compared to those in the lowest risk quantile.[4,6,24] To extend this approach to two separate omics risk scores is more complex, as fewer subdivisions lead to larger group sizes and more statistical power, but more subdivisions allow a more extreme comparison group. As genetic and transcriptomic data were used to define the subtypes, proteomics would provide a third "view" of the data. Thus, to determine the optimal quantiles we split the PRS and TRS into 2 to 3 quantiles (minimum of 4 and maximum of 9 groups) and tested to see what group divisions maximised the number of associated differentially expressed proteins using limma,[18] comparing each quantile category to the lowest quantile. To test the sensitivity of the groups to partitioning, we also examined clinical characteristics for each combination of these partitions. Benjamini-Hochberg[25] false discovery rate (FDR)-adjusted p-values less than 0.05 were considered significant. The number of significantly differentially expressed proteins associated with each omics risk score category was summed.

Clinical comparisons of omics-defined subtypes

We compared clinical characteristics across omics-defined subtypes using the tableone R package. In COPDGene, transcriptomic and proteomic data were collected at the 5-year follow up visit (i.e., "Phase 2"), so we examined differences in anthropometry (including change in BMI per year (Kg/m^2/year) from enrolment to the 5-year follow up visit), spirometry (including prospective FEV_1 change (from 5- to 10-year follow up visits)), and CT measures of emphysema (quantitative emphysema on inspiratory CT scans (% LAA < –950 HU),[26] 15th percentile of lung density histogram on inspiratory CT scans (Perc15)[27]) and of airway thickening (wall area percent (WA%)[26] and square root of wall area of a hypothetical internal perimeter of 10 mm (Pi10)[28]) at the 5-year follow up visit. In ECLIPSE, we examined the same outcomes but longitudinal follow up was from the time of study enrolment to the 3-year follow up visit. In both cohorts, we used post-bronchodilator spirometry measures. We adjusted regression models for age, sex, current smoking status, pack-years of smoking history, principal components of genetic ancestry, and CT scanner (for imaging outcomes only). Sex and race were determined by self-report and checked for concordance with respect to sex and ancestry as assessed by X and Y chromosome and ancestry principal components, as performed in prior genetic association studies.[20,29] To test for an interaction between the PRS and TRS, we tested the significance of the cross-product term (PRS X TRS) in multivariable models adjusting for confounders.

We defined FEV_1 change in COPDGene as the 10-year follow-up measure minus the 5-year follow-up measure, divided by the time in years, and in ECLIPSE, by taking the slope of the best fit line for FEV_1 versus time, as previously described.[7,30] In COPDGene, we additionally examined differences in early-onset COPD (GOLD 2–4 spirometry grades before 55 years of age[31]) and absolute white blood cell differential counts.

As a sensitivity analysis, we included only participants with COPD (FEV1/FVC <0.7) at either baseline or the 5-year follow up visit and repeated the regression analyses.

Regression model specifications. We performed multivariable linear regressions in the COPDGene testing set comparing each subtype to the reference group. Outcomes included FEV_1% predicted, FEV_1/FVC, % LAA < –950 HU, Perc15, Pi10, and WA%. We adjusted regression models for age, sex, current smoking status, pack-years of smoking history, principal components of genetic ancestry, and CT scanner (for imaging outcomes only). We selected outcomes based on input from clinicians and data availability (less than 20% missingness). Adjustment variables were chosen based on clinician input and these measures were available for all participants. We additionally performed interaction analyses by including the PRS, TRS, and a cross-product term (PRS*TRS) within a single regression model.

Biological characterisation of omics-defined subtypes

We performed differential gene and protein expression analyses (accepting FDR-adjusted p-values <0.05), comparing high risk subtypes to the reference group, defined as the group with the lowest PRS and TRS quantiles. We mapped differentially expressed proteins to the human protein–protein interactome[32] and performed Reactome[33] pathway enrichment, Enrichr,[34–36] and STRING[37] analyses.

Differential expression analysis. We performed differential protein expression analysis to identify proteins associated with the PRS, using limma and considering FDR-adjusted p-values below 0.05 to be significant. We also compared differential protein expression between the high-risk groups. To understand how the PRS modifies gene expression profiles, we also examined differentially expressed genes associated with COPD case–control status (GOLD 2–4 versus normal spirometry) adjusting for age, sex, smoking status, pack-years of smoking. We then repeated this analysis adjusting for the PRS and principal components of genetic ancestry. We chose adjustment variables based on clinician input and the fact that these measures were available for all participants.

STRING, pathway enrichment, and drug repurposing analyses. We used STRING (www.string-db.org) to query the human protein–protein interactome and to construct a network including up to 10 interactors in the first shell (i.e., proteins directly interacting with seed proteins) and 5 interactors in the second shell (i.e., proteins directly interacting with 1st shell proteins) per seed protein[37]; only high confidence interactions (≥0.7) were included.

We also performed pathway enrichment and MCL (Markov clustering algorithm, inflation parameter 3) clustering analyses[38] on these protein–protein interaction (PPI) networks. We input the differentially expressed proteins into Enrichr (maayanlab.cloud/Enrichr) to query the Multi-marker Analysis of GenoMic Annotation (MAGMA) drugs and diseases database,[39] and used the Enrichr Appyter to identify potential drug repurposing candidates for individuals belonging to omics-defined subtypes. After these analyses, we renamed the subtype groups based on associated clinical outcomes and biological processes.

Ethics

All study participants provided written informed consent, and studies were approved by local Institutional Review Boards. The current study was approved by the Mass General Brigham institutional review board (IRB #2007P000554).

Role of funders

For ECLIPSE, GlaxoSmithKline was involved in the study design and genotype and phenotype data collection. Otherwise, the study design, data collection, data analysis, data interpretation, and manuscript writing did not involve any sponsors. The final responsibility to submit the publication fell upon the corresponding author, who had full access to all data.

Results

Characteristics of study populations

We included 3274 participants across two cohorts of individuals who smoked. The COPDGene training and testing sets are similar in demographic and spirometry characteristics (Table 1). A diagram of the included and excluded participants is shown in Figure S2. Compared to COPDGene, ECLIPSE participants were more likely to be younger, male, to have a greater number of smoking pack-years, a lower FEV_1% predicted, and lower FEV_1/FVC, and were less likely to be current smokers.

Defining polygenic and transcriptional risk score divisions

Participants plotted on the axes of PRS and TRS exist along a continuum (Figure S1), as is the case for spirometric measures of COPD severity (i.e., FEV_1 and FEV_1/FVC[10,40]). We calculated the gap statistic over a

Characteristic	COPDGene training set	COPDGene testing set (NHW)	COPDGene testing set (AA)	ECLIPSE	p
n	1374	1133	299	468	
Age in years (mean (SD))	67.38 (8.25)	68.13 (8.30)	61.01 (6.95)	64.43 (6.09)	<0.001
Sex (No. (% female))	677 (49.3)	560 (49.4)	153 (51.2)	156 (33.3)	<0.001
Pack-years of smoking (mean (SD))	45.47 (24.61)	45.47 (23.65)	39.55 (20.74)	49.33 (26.87)	<0.001
Current smoking (No. (%))	343 (25.0)	289 (25.5)	182 (60.9)	70 (15.0)	<0.001
FEV1% predicted (mean (SD))	78.37 (24.21)	77.85 (24.81)	81.79 (23.54)	44.22 (14.64)	<0.001
FVC % predicted (mean (SD))	87.25 (17.34)	86.79 (17.55)	88.68 (16.63)	80.63 (19.37)	<0.001
FEV1/FVC (mean (SD))	0.67 (0.15)	0.66 (0.15)	0.71 (0.14)	0.45 (0.11)	<0.001

COPD, chronic obstructive pulmonary disease. COPDGene, Genetic Epidemiology of COPD. ECLIPSE, Evaluation of COPD to Longitudinally Identify Predictive Surrogate Endpoints study. FEV_1, forced expiratory volume in 1 s. FEV_1/FVC, FEV1/forced vital capacity. NHW, non-Hispanic white. AA, African American.

Table 1: **Characteristics of study populations.**

range of kmeans cluster numbers in the COPDGene training set, and consistent with visual inspection of Figure S1, we observed that one cluster yields the highest gap statistic, indicating that there are no clusters (Figure S3). As an alternative approach to clustering, we applied the common practice of dividing risk scores into quantiles, though the optimal quantiles balancing sufficiently high risk, yet adequate sample size, are not clear. Thus, we tested four combinations of partitioned omics risk scores, using protein expression differences (not used in the PRS and TRS) between groups. We observed that dichotomising PRS and dividing TRS into tertiles yielded the greatest number of differentially expressed proteins (Table S1). We then applied these same quantiles to the COPDGene testing set and ECLIPSE participants (Fig. 2). To test the robustness of subgroups to specific partitions, we also examined 4 to 9 subdivisions and noted stable clinical characteristics of the highest ("low PRS, high TRS" and "high PRS, high TRS") and lowest risk ("low PRS, low TRS") groups (Table S2).

Polygenic and transcriptional risk scores identify "high disease activity" and "severe disease risk" subtypes

We observed, as expected, heterogeneity amongst the two high-risk (i.e., high TRS) groups (Table 2). We compared these high-risk subtypes to a reference group, which was defined as the lowest omics risk group (i.e., "low PRS, low TRS" subtype). Compared to the reference group, the two high-TRS risk groups (i.e., "low PRS, high TRS" and "high PRS, high TRS") demonstrated decreased BMI, lower spirometry measures, more emphysema, and thicker airways across testing cohorts (Table 2). Both groups had similar mean adjusted prospective FEV_1 decline in the COPDGene testing set compared to the reference group, but this finding was only consistent for the "low PRS, high TRS" group in ECLIPSE (–40 mL/year).

We then performed linear regression analyses on selected COPD-related outcomes. We compared anthropometric, spirometry, CT, and other COPD-related outcomes across COPDGene and ECLIPSE (Table 3, Table S3). Compared to the reference group,

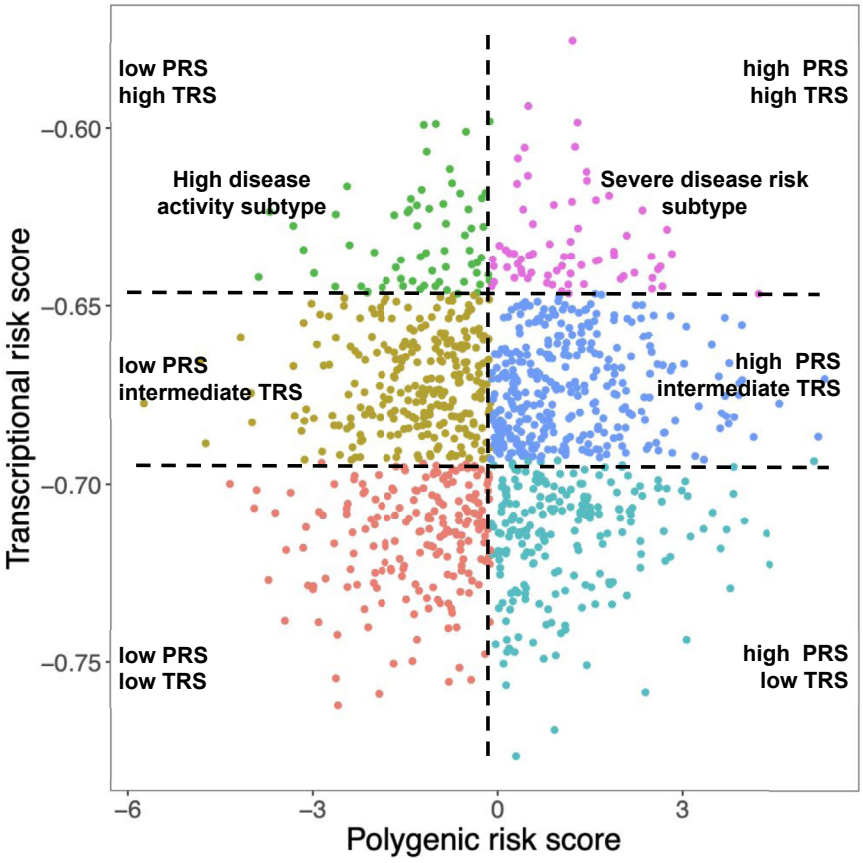

Fig. 2: Omics-defined groups or subtypes overlaid on a plot of the polygenic risk score (PRS; *x-axis*) and transcriptional risk score (TRS; *y-axis*) in the COPDGene testing set (n = 1432).

Characteristic	COPDGene NHW testing set			COPDGene AA testing set			ECLIPSE		
	Reference group (Low PRS, Low TRS)	Low PRS, High TRS	High PRS, High TRS	Reference group (Low PRS, Low TRS)	Low PRS, High TRS	High PRS, High TRS	Reference group (Low PRS, Low TRS)	Low PRS, High TRS	High PRS, High TRS
n	196	65	68	61	16	11	57	63	56
Age in years (mean (SD))	67.85 (8.52)	67.16 (6.46)	69.04 (8.11)	61.26 (7.55)	60.58 (7.24)	65.95 (8.05)	63.00 (5.10)	65.43 (6.65)	66.59 (5.95)
Sex (No. (% female))	114 (58.2)	19 (29.2)	24 (35.3)	36 (59.0)	5 (31.2)	8 (72.7)	30 (52.6)	10 (15.9)	11 (19.6)
Pack-years of smoking (mean (SD))	36.51 (22.27)	60.07 (26.49)	57.28 (23.35)	37.01 (20.85)	43.42 (16.14)	61.36 (38.87)	45.49 (26.24)	55.29 (35.25)	54.29 (21.10)
Current smoking (No. (%))	20 (10.2)	36 (55.4)	36 (52.9)	31 (50.8)	15 (93.8)	8 (72.7)	6 (10.5)	7 (11.1)	12 (21.4)
BMI (Kg/m^2) (mean (SD))	30.00 (6.84)	26.51 (5.81)	25.70 (5.58)	30.61 (6.36)	29.14 (8.52)	25.26 (6.57)	27.54 (5.06)	26.70 (6.22)	25.49 (4.76)
FEV1% predicted (mean (SD))	90.15 (18.13)	66.50 (27.07)	60.55 (24.11)	90.83 (17.41)	80.61 (25.43)	71.08 (10.67)	50.25 (13.42)	43.41 (15.65)	37.44 (13.44)
FVC % predicted (mean (SD))	90.44 (15.27)	85.25 (20.18)	83.25 (16.65)	91.65 (14.37)	85.12 (20.95)	91.99 (13.77)	90.53 (20.17)	79.60 (18.68)	75.25 (22.51)
FEV1/FVC (mean (SD))	0.75 (0.09)	0.57 (0.15)	0.53 (0.15)	0.77 (0.10)	0.73 (0.14)	0.61 (0.12)	45.84 (11.13)	43.16 (12.79)	39.71 (10.29)
% LAA < −950 HU (mean (SD))	2.78 (4.71)	8.50 (10.39)	11.00 (11.98)	1.71 (4.00)	3.51 (8.08)	5.49 (7.30)	16.01 (11.46)	20.88 (13.50)	22.44 (13.22)
Perc15 (mean (SD))	−913.62 (20.43)	−927.23 (27.21)	−931.05 (31.55)	−895.53 (26.89)	−901.19 (29.94)	−913.62 (34.95)	−943.75 (50.13)	−960.00 (49.44)	−971.11 (45.84)
Pi10 (mean (SD))	2.01 (0.50)	2.55 (0.65)	2.60 (0.51)	2.11 (0.48)	2.56 (0.59)	2.42 (0.44)	4.38 (0.20)	4.42 (0.23)	4.43 (0.19)
WA % (mean (SD))	46.97 (8.11)	54.46 (8.35)	54.61 (7.41)	47.48 (7.76)	54.86 (10.52)	51.62 (4.06)	64.35 (2.79)	65.05 (3.58)	67.22 (3.44)
Change in FEV1 mL/year (prospective) (mean (SD))	−32.11 (48.07)	−51.21 (74.41)	−66.45 (61.58)	−42.28 (36.59)	−87.86 (156.98)	−56.26 (76.62)	−32.21 (59.54)	−40.13 (75.31)	−15.22 (67.46)

The reference group was defined as the "Low PRS, Low TRS" group. BMI, body-mass index. LAA, low attenuation area. HU, Hounsfield units. Perc15, 15th percentile of lung density histogram on inspiratory CT scans. WA %, wall area percent. Pi10, square root of wall area of a hypothetical internal perimeter of 10 mm. ACO, asthma-COPD overlap. See Table 1 legend for other abbreviations. NHW, non-Hispanic white. AA, African American.

Table 2: **Omics-defined subtypes in the COPDGene testing set and ECLIPSE.**

high-risk groups exhibited lower spirometry, more emphysema, and thicker airways (Table 3). While none of the adjusted FEV_1 decline measures were statistically significant, the "low PRS, high TRS" means were consistent between COPDGene and ECLIPSE (−30 mL/yr and −24 mL/year, p = 0.15 and p = 0.11, respectively), despite that COPDGene participants had only two FEV_1 measurements while ECLIPSE participants had up to six FEV_1 measurements. Given the clinical relevance of accelerated FEV_1 decline in the "low PRS, high TRS" group, we renamed this group the "high disease activity" subtype. While the "high PRS, high TRS" group did not have replicable FEV_1 decline across cohorts, this group had the lowest lung function and most emphysema; therefore, we renamed this group the "severe disease risk" subtype. We observed similar distributions of COPD-related outcomes across omics-defined subtypes when stratifying by sex (Table S4). In COPDGene NHW participants only, the "high disease activity" subtype also exhibited a trend toward greater decline in BMI compared to the reference group (β = −0.154 [95% CI: −0.333 to 0.0253], p = 0.094). In multivariable interaction analyses, we did not observe any significant interactions (all p-values of cross-product terms [PRS X TRS] > 0.05) between the PRS and TRS on spirometry or CT imaging outcomes (Table S5). As a sensitivity analysis, we limited analysis to those with COPD (FEV1/FVC <0.7) at either study visit (Table S6). We note that all ECLIPSE individuals had COPD. The results in COPDGene are similar with respect to characteristics across omics-defined subtypes (Table S7).

Biological characterisation and drug repurposing analyses of subtypes

Having identified clinical differences between the two high-risk groups, we sought to characterise biological differences between these subtypes. We performed differential gene and protein expression analyses between

COPD-related outcome	COPDGene NHW testing set				COPDGene AA testing set				ECLIPSE			
	Low PRS, High TRS		High PRS, High TRS		Low PRS, High TRS		High PRS, High TRS		Low PRS, High TRS		High PRS, High TRS	
	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p	beta (95% CI)	p
FEV1% predicted	−22.3 (−29.5 to −15.2)	4.3E-09	−31.5 (−38.2 to −24.8)	1E-17	−4.11 (−16.8 to 8.59)	0.53	−19.6 (−32.7 to −6.51)	0.0048	−7.8 (−13.6 to −1.95)	0.01	−15.5 (−21.4 to −9.61)	1.3E-06
FVC % predicted	−7.24 (−13 to −1.51)	0.014	−11.9 (−17.1 to −6.63)	1.3E-05	−4.07 (−14.6 to 6.45)	0.45	−3.77 (−15.5 to 7.91)	0.53	−6.89 (−14.9 to 1.15)	0.096	−12.4 (−22.2 to −2.51)	0.016
FEV1/FVC	−0.147 (−0.184 to −0.11)	3E-13	−0.197 (−0.233 to −0.161)	4.8E-22	−0.00368 (−0.0731 to 0.0657)	0.92	−0.118 (−0.193 to −0.043)	0.0032	−4.5 (−9.07 to 0.0772)	0.057	−8.83 (−13.5 to −4.16)	0.00035
% LAA < −950 HU	3.1 (1.8–5.2)	3.30E-05	3.4 (2.1–5.6)	2.30E-06	1.3 (0.44–3.7)	0.66	0.94 (0.23–3.9)	0.93	2.33 (−3.34 to 8)	0.42	6.97 (−0.0748 to 14)	0.058
Perc15	−11.2 (−18.6 to −3.72)	0.0037	−15.3 (−22.8 to −7.89)	7.6E-05	−0.305 (−17.1 to 16.5)	0.97	−4.66 (−26.2 to 16.9)	0.67	−0.994 (−24.1 to 22.1)	0.93	−19.2 (−49.3 to 10.8)	0.21
Pi10	0.498 (0.298–0.698)	2.10E-06	0.605 (0.429–0.78)	1.40E-10	0.346 (−0.032 to 0.724)	0.079	0.475 (0.00489–0.946)	0.054	−0.0786 (−0.147 to −0.00988)	0.029	−0.0753 (−0.163 to 0.0124)	0.099
WA %	5.88 (3.19–8.58)	2.90E-05	7.17 (4.63–9.71)	8.90E-08	4.55 (−1.6 to 10.7)	0.15	7.48 (−0.0534 to 15)	0.058	0.516 (−0.972 to 2)	0.5	2.18 (0.376–3.98)	0.021
FEV1 change (mL/year)	−30 (−70.6 to 10.7)	0.15	−43.9 (−91.7 to 3.94)	0.077	−103 (−294 to 87.5)	0.32	−60.5 (−105 to −15.6)	0.033	−23.7 (−52.9 to 5.47)	0.11	6.97 (−25.8 to 39.8)	0.68

Models were adjusted or age, sex, current smoking status, pack-years of smoking, and principal components of genetic ancestry. Computed tomography imaging outcomes were additionally adjusted for CT scanner. Abbreviations are detailed in Tables 1 and 2 legends.

Table 3: **Multivariable linear regressions in the COPDGene testing sets and ECLIPSE.**

the "high disease activity" and "severe disease risk" subtypes and the reference group in the COPDGene NHW testing set (Tables S8–S10). The "high disease activity" subtype had 14 and the "severe disease risk" subtype had 2 differentially expressed proteins (Table S8). We did not observe differentially expressed genes or proteins when directly comparing high risk groups. We examined how the PRS affects differential gene expression associated with COPD case–control status as detailed in the supplement (Supplementary Results and Table S11).

We mapped differentially expressed proteins associated with each high-risk subtype in the COPDGene testing set to the human protein–protein interactome[32] and used the mapped proteins as seed proteins to construct STRING PPI networks (Figs. 3 and 4) with associated MCL clusters (Table S12) and perform pathway enrichment analyses (Table S13). We built separate up and down-regulated STRING PPI networks for the low PRS/high TRS group, and additionally observed alterations in the Asparagine N-linked glycosylation, NCAM1, and RAF/MAP kinase cascade pathways (Table S14). To identify subtype-specific drug repurposing candidates, we used these same seed proteins to perform enrichment analyses on the MAGMA Drugs and Disease database.[39] Both subtypes demonstrated enrichment proteomic profiles suggesting potential treatment with ACE inhibitors, thyroid medications, carvedilol, bromocriptine, and lovastatin; the "high disease activity" subtype also had significant findings for 5-lipoxygenase inhibitors, fomepizole, and galantamine, while the "severe disease risk" (Table S15) subtype had significant findings for atypical antipsychotics.

Discussion

In this study of 3274 ever-smokers from two cohorts, we used blood-based polygenic (PRS) and transcriptional (TRS) risk scores to identify heterogeneity within high-

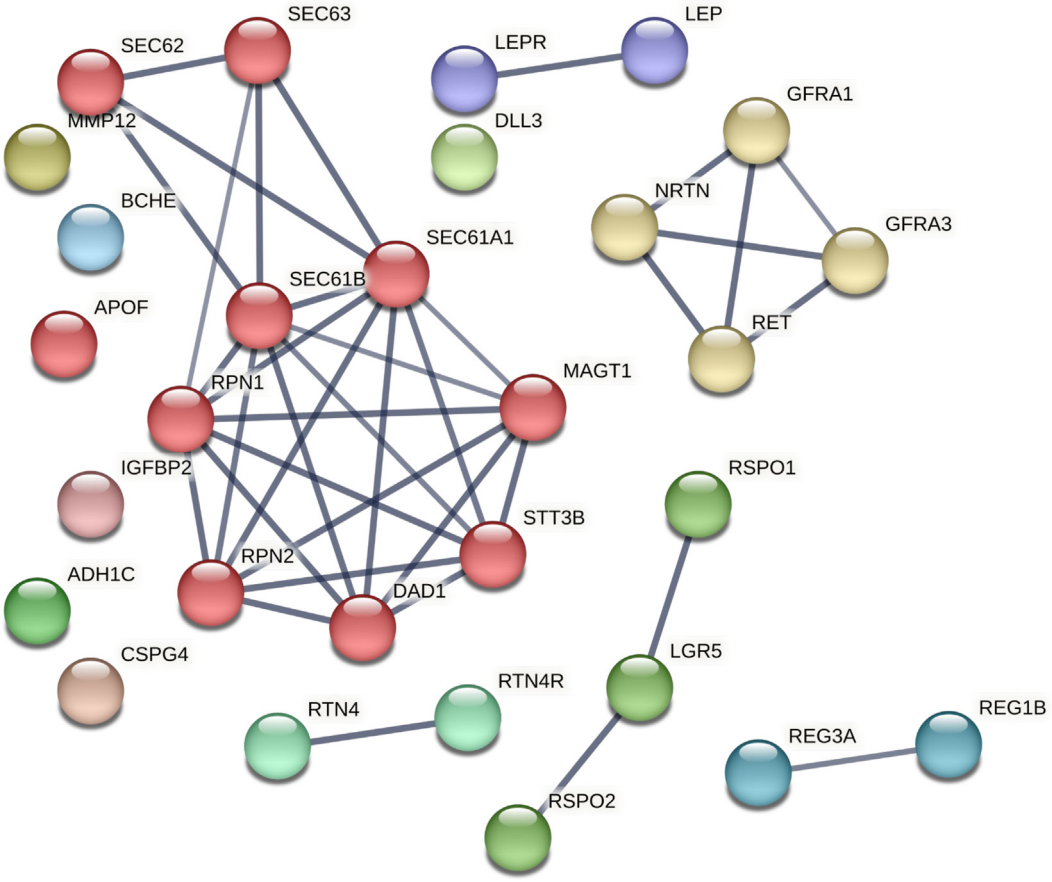

Fig. 3: High disease activity ("low PRS, high TRS") subtype STRING protein–protein interaction networks using differentially expressed proteins in Omics-defined groups (subtypes) in the COPDGene testing set as seed proteins, permitting up to 10 interactors in the first shell and 5 interactors in the second shell. Only high-confidence interactions were included and greater line thickness indicates greater confidence. Differentially expressed proteins were identified by comparing group assignments to the reference group. Colours represent MCL (Markov) clusters.

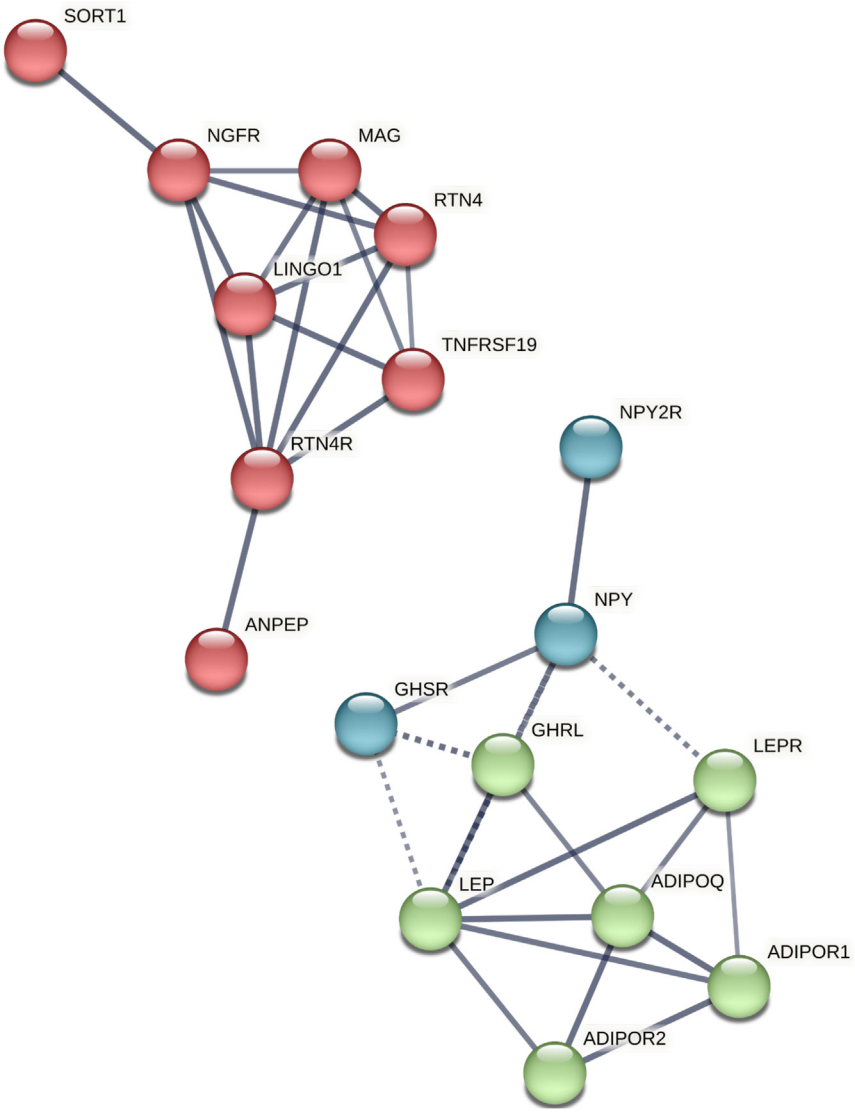

Fig. 4: Severe disease risk ("high PRS, high TRS") subtype STRING protein–protein interaction networks using differentially expressed proteins in Omics-defined groups (subtypes) in the COPDGene testing set as seed proteins, permitting up to 10 interactors in the first shell and 5 interactors in the second shell. Only high-confidence interactions were included and greater line thickness indicates greater confidence. Differentially expressed proteins were identified by comparing group assignments to the reference group. Colours represent MCL (Markov) clusters.

risk individuals, defining "high disease activity" and "severe disease risk" COPD subtypes. Compared to a reference group, both subtypes had lower mean BMI values and alterations in metabolic, growth, and immune signalling processes. "High disease activity" participants exhibited prospective FEV_1 decline across both replication cohorts, albeit with a non-significant (though directionally consistent) association in multivariable models adjusted for baseline FEV_1. "Severe disease risk" participants had low spirometry measures with high quantitative emphysema and thick airways. We identified biological processes and drug repurposing candidates associated with each subtype, including therapies previously tested in COPD clinical trials. Our study demonstrates how omics risk scores can identify COPD subtypes with associated clinical and biological characteristics that can be leveraged for therapeutic interventions.

Linking omics-defined high-risk groups to specific pathobiology is an active area of research that we now extend to lung disease. In schizophrenia, PRS-defined high risk groups were biologically characterised using

weighted gene co-expression network analyses.[41] Other approaches have incorporated gene expression into PRSs to improve prediction and imply biological mechanisms.[42–44] Here, we used genetic and transcriptomic data to define subtypes, and then leveraged proteomic differences between groups to understand subtype biology. Importantly, our results suggest that different omics risk scores are not interchangeable, i.e., a higher omics risk score will not always have the same association with specific outcomes. Individuals with the highest transcriptomic quantile exhibited different clinical and biological features depending on their underlying polygenic risk. We did not observe statistically significant interactions between the PRS and TRS on COPD-related outcomes in regression analyses, but our quantile-based omics-defined subtyping approach identified individuals with severe disease and divergent clinical characteristics; our approach was able to identify heterogeneity that was not detected using traditional regression-based interaction analyses. We also demonstrate that individuals may exist along a continuum of COPD risk—i.e., there are no clusters—yet omics-defined subgroups may have important clinical and biological differences. While we did not observe clusters, we observed that omics-defined subgroups were robust to varying PRS and TRS subdivisions, with the two high-risk groups ("low PRS, high TRS" and "high PRS, high TRS") demonstrating stable clinical characteristics across risk score subdivisions. Thus, genetics and transcriptomics may provide alternative yet complementary views of lung function biology.

The two high-risk groups we identified have important clinical differences. Participants in these groups have different degrees of lung function impairment, amounts of emphysema, airway wall pathology, and lung function decline. For clinical prognosis, it is helpful to identify which patients are more severely affected and will decline quickly. Using blood-based omics alone, we were able to identify subtypes of individuals with very advanced disease (high PRS, high TRS) as well as those prone to accelerated FEV_1 decline (low PRS, high TRS).

Of direct clinical relevance, the association of lower spirometry, greater emphysema and observed FEV_1 decline across two cohorts suggests that the "high disease activity" subtype is a targetable trait, the risk of which might be modified by approved medications (5-lipoxygenase inhibitors, angiotensin-converting enzyme (ACE) inhibitors, fomepizole, galantamine). Fomepizole has not previously been implicated as a COPD drug-repurposing candidate, to our knowledge. Galantamine is known to cause bronchospasm, and enrichment for proteins targeted by galantamine suggests this drug is most likely to cause harm in this subtype of patients. Although a randomised trial of the 5-lipoxygenase inhibitor, Zileuton, did not reduce length of stay or treatment failure in patients hospitalised for COPD exacerbations, it was likely underpowered,[45] and did not examine longer term outcomes. ACE inhibitors and angiotensin receptor blockers (ARB) have been identified as COPD drug repurposing candidates. A recent clinical trial showed failure of losartan to decrease emphysema progression.[46] Our drug repurposing analyses implicated utility of captopril—not all ACE inhibitors and ARBs–in the "high disease activity" but not the "severe disease risk" subtype. Conversely, our analysis suggests potential benefit of atypical antipsychotics in the "severe disease risk" subtype. The coincidence of schizophrenia and COPD is largely attributed to smoking, though a phenome-wide association and polygenic risk analysis suggests that schizophrenia and obstructive lung disease may have shared genetic mechanisms.[47] While we adjusted for self-reported cigarette smoking status, these measures are imperfect, and thus we cannot address whether shared mechanisms are due to smoking behaviour. A broader implication of the drug repurposing analyses that merits validation is that previous failure of drugs in clinical trials was due to heterogeneity in patient selection that could be overcome by omics-based subtyping.

The cachectic COPD patient, who may be prone to more exacerbations, is a well-described clinical phenotype,[48–51] and we observed that the "high disease activity" and "severe disease risk" subtypes have lower mean BMI than other subtypes. Although pulmonary cachexia has been proposed to occur only in severe COPD,[52] we demonstrate additional heterogeneity within lower BMI ever-smokers and identify a "high disease activity" subtype with less spirometric severity and distinct biology. The "high disease activity" group may also overlap with a previously identified comorbidity cachexia subgroup,[53] and in the current analysis, we identified molecular profiles that might guide therapy. Relevant to body composition, the adipocyte product leptin exhibited altered expression in both subtypes, which has several potential, not mutually exclusive interpretations. Leptin acts both as a hormone negatively regulating hunger and adipocyte fat storage[54] and as a proinflammatory cytokine essential for host defence.[55,56]

The role of leptin in regulating hunger suggests it could be relevant to the "severe disease risk" subtype, which had the largest effect size in leptin expression (–0.71 log-fold change); the observed decrease in leptin could be a compensatory response to or share a causal relationship with pulmonary cachexia. In addition, the lower leptin levels observed in malnourished populations are associated with dysfunctional cell-mediated immunity and increased susceptibility to infections.[55,57] In COPD, elevated leptin concentrations have been reported in the plasma and airways in some,[58–62] but not all studies,[63] and has been identified as a potential biomarker of emphysema progression.[64,65] Our results suggest that "high disease activity" or "severe disease risk" subtype individuals might exhibit humoral and

cytokine profiles similar to those seen in malnourished individuals with increased susceptibility to infections. It remains unclear whether the observed peripheral blood proteomic alterations are merely a consequence of disease activity, or a causal component of a positive feedback loop involved in driving disease progression. The biomarkers used in this study are blood-based, making it difficult to identify the relevant mechanisms in lung tissue. However, blood-based biomarkers are practical in a clinical setting and follow up studies linking changes in these biomarkers to pathophysiologic mechanism in lung tissue can help to bridge the gap between prediction and precision therapeutics.

This study leveraged multiple omics data in the form of validated, replicated, risk scores, to identify COPD subtypes in two cohorts of ever-smokers. Often, the poor correlation between transcripts and proteins precludes them from being used within the same analyses, but we were able to leverage the weak correlation to understand protein-level biological changes occurring in subtypes defined using genetics and transcriptomics. The proteome, which we used for protein–protein interaction analyses, provides several advantages in defining biological subtypes in omics studies. Proteins are the functional molecules in cells, directly involved in biological processes, pathways, and interactions, and tend to provide easily attainable biomarkers. As many existing drugs are small molecules that target proteins, utilising proteomic data is preferrable for drug target identification and repurposing analyses.

One major challenge of this analysis was determining how to subset subjects from a standardised, continuous distribution. To overcome this limitation, we varied the number of quantiles and assessed which combination of divisions yielded the greatest number of differentially expressed proteins, then identified corresponding raw score cut-off values that allowed each participant in the COPDGene and ECLIPSE testing sets to be categorised into an omics-defined subtype. We acknowledge that there are other reasonable approaches for applying omics risk scores to identify COPD subtypes. We also note that COPDGene and ECLIPSE are cohorts that represent heavy smokers, and so our approach applies to patients meeting this study inclusion criteria.

Individuals near the cut-off value for PRS might exist on a continuum between "high disease activity" and "severe disease risk" subtypes, and these participants might be able to transition between subtypes; regardless, defining thresholds and categories are an important step toward clinical translation. We focused on the highest risk groups; while the high PRS, low TRS group would theoretically be of interest, we observed that participants in this group had comparatively mild disease compared to the other severe groups. The FEV_1/FVC cutoff of 0.7 for airflow obstruction has come into question with respect to disease definitions. Indeed, recent ATS/ERS spirometry guidelines[66] now recognise preserved ratio with impaired spirometry (PRISm) individuals who experience respiratory symptoms and have emphysema, airway wall pathology, increased mortality, and exacerbations.[40,67–69] However, there is also an upward inflection point in mortality as the FEV_1/FVC falls below 0.7.[70] Therefore, we repeated our main analyses including only those with FEV_1/FVC less than 0.7 and saw similar results, suggesting that this omic-based subtyping approach applies to patients with COPD as well as smokers at high risk for disease.

We observed similar results in COPDGene AA participants, including the "high disease activity" subtype association with FEV_1 decline in multivariable analyses. We previously described how the PRS, developed in European ancestry individuals, had decreased performance in individuals of non-European ancestry.[6] While we tested our subtypes in African Americans, we emphasise that future studies are needed to improve genetic prediction in non-Europeans. While we used COPDGene as a discovery cohort and ECLIPSE for validation, the cohorts have notable differences. COPDGene included a larger number of participants without COPD, while ECLIPSE included a larger proportion of severe COPD. Additionally, ECLIPSE used a lower-dose CT scan. We have previously observed disparate measures of emphysema and airway wall thickness even after propensity score-matching participants from these two cohorts,[71] which is likely related to differences in imaging protocols. We based our drug repurposing candidates on enrichment analyses of proteomic profiles, but directionality of biological processes was not accounted for in enrichment analyses; additional mechanistic and clinical trial validations are needed. Finally, we identified two high-risk subtypes using this approach, but these do not explain all of COPD heterogeneity, and there are almost certainly other important subtypes.

In conclusion, polygenic and transcriptional risk scores, both based on spirometry, identified "high disease activity" and "severe disease risk" subtypes with distinct clinical and biological characteristics. Proteomic and drug repurposing analysis identified subtype-specific enrichment for therapies, some of which were previously hypothesised in COPD.

Contributors

All authors read and approved the final version of the manuscript. The data were accessed and verified by Michael H. Cho, Brian D. Hobbs, and Matthew Moll.

Study Design: Matthew Moll, Julian Hecker, Michael H. Cho, Brian D. Hobbs.

Acquisition, analysis, or interpretation of the data: Matthew Moll, Julian Hecker, Edwin K. Silverman, Auyon Ghosh, Don D. Sin, Peter J. Castaldi, Katherine Pratte, Russell Bowler, Gary Anderson, Brian D. Hobbs, Michael H. Cho.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Matthew Moll, Julian Hecker, John Platig, Kimberly Glass, Brian D. Hobbs, Michael H. Cho.

Obtained funding: Edwin K. Silverman, Michael H. Cho, Matthew Moll.

Data sharing statement

Genetic, transcriptomic, and proteomic data are publicly available through dbGaP (COPDGene: phs000179.v1.p1; ECLIPSE: phs001252.v1.p1). R code can be made available upon request at remol@channing.harvard.edu.

Declaration of interests

EKS received grant support from Bayer and Northpond Labs. BDH received grant support from Bayer. MHC has received grant support from Bayer. MM received grant support from Bayer and consulting fees from Sitka, TheaHealth, 2ndMD, and TriNetX. CPH reports grant support from Boehringer-Ingelheim, Novartis, Bayer and Vertex, outside of this study. PJC has received grant support from GlaxoSmithKline and Bayer and consulting fees from GlaxoSmithKline and Novartis. RTS received consulting fees from GSK, AstraZeneca, Roche, Itai and Beyond, Samay Health, Immunomet, ENA Respiratory, Teva, COPD Foundation and Vocalis Health. She is a retiree and shareholder of GSK and holds share options at ENA Respiratory. DDS received honoraria for giving talks on COPD from GSK, Boehringer Ingelheim, and AstraZeneca, is the chair of an NHLBI sponsored clinical trial data safety monitoring board, and deputy editor of European Respiratory Journal. JLC received consulting fees from AstraZeneca PLC, CSL Behring, LLC, and Novartis Corporation. SIR received consulting fees from Verona Pharma, Sanofi, BeyondAir and the Alpha 1 Foundation. He is a founder and president of Great Plains Biometrix. He was an employee of AstraZeneca from 2015 to 2019 during which he received shares as part of his compensation. JHP is supported by NIH K25HL140186. KG is supported by NIH/NHLBI: R01HG011393, R01HL152728, R01HL160008, and R01HL162813.

Acknowledgements

MM is supported by K08HL159318.

BDH is supported by NIH K08HL136928, U01 HL089856, and an Alpha-1 Foundation Research Grant.

JH is supported by P01 HL132825.

MLM is supported by NIH R01HL153460 and VA 1I01RX002745.

CPH is supported by NIH R01HL157879 and P01HL114501.

MHC is supported by NIH R01HL137927, R01HL135142, HL147148, and HL089856.

PJC is supported by NIH R01HL124233 and R01HL147326.

RPB is supported by NIH R01 HL137995 and R01 HL152735.

EKS is supported by NIH R01 HL147148, U01 HL089856, R01 HL133135, R01 HL152728, and P01 HL114501.

GPA is supported by the Australian federal government via grants from NHMRC of Australia and Medical Research Futures Fund.

Proteomic data generated for this proposal was supported by R01 HL137995.

The COPDGene project was supported by NHLBI grants U01 HL089897 and U01 HL089856 and by NIH contract 75N92023D00011. 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. COPDGene is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion. The ECLIPSE study (NCT00292552; GSK code SCO104960) was funded by GlaxoSmithKline.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2024.105429.

References

1 Safiri S, Carson-Chahhoud K, Noori M, et al. Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the Global Burden of Disease Study 2019. *BMJ*. 2022;378:e069679.
2 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.
3 Wedzicha JA. The heterogeneity of chronic obstructive pulmonary disease. *Thorax*. 2000;55(8):631–632.
4 Natarajan P, Young R, Stitziel NO, et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*. 2017;135(22):2091–2101.
5 Rodon J, Soria JC, Berger R, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nat Med*. 2019;25(5):751–758.
6 Moll M, Sakornsakolpat P, Shrine N, et al. Chronic obstructive pulmonary disease and related phenotypes: polygenic risk scores in population-based and case-control cohorts. *Lancet Respir Med*. 2020;8(7):696–708.
7 Moll M, Boueiz A, Ghosh AJ, et al. Development of a blood-based transcriptional risk score for chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2022;205(2):161–170.
8 Zhang J, Xu H, Qiao D, et al. A polygenic risk score and age of diagnosis of chronic obstructive pulmonary disease. *Eur Respir J*. 2022:2101954.
9 Sun J, Wang Y, Folkersen L, et al. Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction. *Nat Commun*. 2021;12(1):5276.
10 Castaldi PJ, Dy J, Ross J, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax*. 2014;69(5):415–422.
11 Martinez FJ, Agusti A, Celli BR, et al. Treatment trials in young patients with chronic obstructive pulmonary disease and pre-chronic obstructive pulmonary disease patients: time to move forward. *Am J Respir Crit Care Med*. 2022;205(3):275–287.
12 Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2010;7(1):32–43.
13 Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype reference consortium panel. *Nat Genet*. 2016;48(11):1443–1448.
14 Vestbo J, Anderson W, Coxson HO, et al. Evaluation of COPD longitudinally to identify predictive surrogate end-points (ECLIPSE). *Eur Respir J*. 2008;31(4):869–873.
15 Global Initiative for Chronic Obstructive Lung Disease - GOLD. 2024 GOLD report [cited 2024 Mar 11] Available from: https://goldcopd.org/2024-gold-report/.
16 Parker MM, Chase RP, Lamb A, et al. Correction to: RNA sequencing identifies novel non-coding RNA and exon-specific effects associated with cigarette smoking. *BMC Med Genom*. 2019;12(1):166.
17 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma Oxf Engl*. 2010;26(1):139–140.
18 Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
19 Serban KA, Pratte KA, Strange C, et al. Unique and shared systemic biomarkers for emphysema in Alpha-1 Antitrypsin deficiency and chronic obstructive pulmonary disease. *eBioMedicine*. 2022;84:104262.
20 Shrine N, Guyatt AL, Erzurumluoglu AM, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet*. 2019;51(3):481–493.
21 Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol*. 2017;41(6):469–480.
22 Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B*. 1996:4196–4249.
23 Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser B Stat Methodol*. 2001;63(2):411–423.
24 Khera AV, Chaffin M, Wade KH, et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*. 2019;177(3):587–596.e9.
25 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289–300.

26 Han MK, Kazerooni EA, Lynch DA, et al. Chronic obstructive pulmonary disease exacerbations in the COPDGene study: associated radiologic phenotypes. *Radiology*. 2011;261(1):274–282.
27 Parr DG, Sevenoaks M, Deng CQ, Stoel BC, Stockley RA. Detection of emphysema progression in alpha 1-antitrypsin deficiency using CT densitometry; Methodological advances. *Respir Res*. 2008;9:1–8.
28 Van Tho N, Ogawa E, Trang LTH, et al. A mixed phenotype of airway wall thickening and emphysema is associated with dyspnea and hospitalization for chronic obstructive pulmonary disease. *Ann Am Thorac Soc*. 2015;12(7):988–996.
29 Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet*. 2019;51(3):494–505.
30 Vestbo J, Edwards LD, Scanlon PD, et al. Changes in forced expiratory volume in 1 second over time in COPD. *N Engl J Med*. 2011;365(13):1184–1192.
31 Foreman MG, Zhang L, Murphy J, et al. Early-onset chronic obstructive pulmonary disease is associated with female sex, maternal factors, and African American race in the COPDGene study. *Am J Respir Crit Care Med*. 2011;184(4):414–420.
32 Wang RS, Loscalzo J. Network module-based drug repositioning for pulmonary arterial hypertension. *CPT Pharmacomet Syst Pharmacol*. 2021;10(9):994–1005.
33 Fabregat A, Sidiropoulos K, Viteri G, et al. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinf*. 2017;18(1):142.
34 Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf*. 2013;14:128.
35 Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. 2016;44(W1):W90–W97.
36 Xie Z, Bailey A, Kuleshov MV, et al. Gene set knowledge discovery with Enrichr. *Curr Protoc*. 2021;1(3):e90.
37 Jensen LJ, Kuhn M, Stark M, et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*. 2009;37(Database issue):D412–D416.
38 Satuluri V, Parthasarathy S, Ucar D. Markov clustering of protein interaction networks with improved balance and scalability. In: *Proceedings of the first ACM international conference on bioinformatics and computational biology*. New York, NY, USA: Association for Computing Machinery; 2010:247–256. https://doi.org/10.1145/1854776.1854812 (BCB '10).
39 de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol*. 2015;11(4):e1004219.
40 Lowe KE, Regan EA, Anzueto A, et al. COPDGene® 2019: redefining the diagnosis of chronic obstructive pulmonary disease. *Chronic Obstr Pulm Dis Miami Fla*. 2019;6(5):384–399.
41 Radulescu E, Jaffe AE, Straub RE, et al. Identification and prioritization of gene sets associated with schizophrenia risk by co-expression network analysis in human brain. *Mol Psychiatry*. 2020;25(4):791–804.
42 Hu X, Qiao D, Kim W, et al. Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program. *Am J Hum Genet*. 2022;109(5):857–870.
43 Mareckova K, Hawco C, Dos Santos FC, et al. Novel polygenic risk score as a translational tool linking depression-related changes in the corticolimbic transcriptome with neural face processing and anhedonic symptoms. *Transl Psychiatry*. 2020;10(1):1–10.
44 Bandres-Ciga S, Saez-Atienzar S, Kim JJ, et al. Large-scale pathway specific polygenic risk and transcriptomic community network analysis identifies novel functional pathways in Parkinson disease. *Acta Neuropathol*. 2020;140(3):341–358.
45 Woodruff PG, Albert RK, Bailey WC, et al. Randomized trial of zileuton for treatment of COPD exacerbations requiring hospitalization. *COPD*. 2011;8(1):21–29.
46 Wise RA, Holbrook JT, Brown RH, et al. Clinical trial of losartan for pulmonary emphysema: pulmonary trials cooperative losartan effects on emphysema progression clinical trial. *Am J Respir Crit Care Med*. 2022;206(7):838–845.
47 Zhang R, Sjölander A, Ploner A, Lu D, Bulik CM, Bergen SE. Novel disease associations with schizophrenia genetic risk revealed in ~400 000 UK Biobank participants. *Mol Psychiatry*. 2022;27(3):1448–1454.
48 McDonald MLN, Wouters EFM, Rutten E, et al. It's more than low BMI: prevalence of cachexia and associated mortality in COPD. *Respir Res*. 2019;20(1):100.
49 Grigsby MR, Siddharthan T, Pollard SL, et al. Low body mass index is associated with higher odds of COPD and lower lung function in low- and middle-income countries. *COPD*. 2019;16(1):58–65.
50 Wada H, Ikeda A, Maruyama K, et al. Low BMI and weight loss aggravate COPD mortality in men, findings from a large prospective cohort: the JACC study. *Sci Rep*. 2021;11(1):1531.
51 Mason SE, Moreta-Martinez R, Labaki WW, et al. Respiratory exacerbations are associated with muscle loss in current and former smokers. *Thorax*. 2021;76(6):554–560.
52 Eriksson B, Backman H, Bossios A, et al. Only severe COPD is associated with being underweight: results from a population survey. *ERJ Open Res*. 2016;2(3):51–2015.
53 Vanfleteren LEGW, Spruit MA, 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.
54 Collins S, Kuhn CM, Petro AE, Swick AG, Chrunyk BA, Surwit RS. Role of leptin in fat regulation. *Nature*. 1996;380(6576):677.
55 La Cava A. Leptin in inflammation and autoimmunity. *Cytokine*. 2017;98:51–58.
56 La Cava A, Matarese G. The weight of leptin in immunity. *Nat Rev Immunol*. 2004;4(5):371–379.
57 Matarese G, La Cava A, Sanna V, et al. Balancing susceptibility to infection and autoimmunity: a role for leptin? *Trends Immunol*. 2002;23(4):182–187.
58 Suzuki M, Makita H, Östling J, et al. Lower leptin/adiponectin ratio and risk of rapid lung function decline in chronic obstructive pulmonary disease. *Ann Am Thorac Soc*. 2014;11(10):1511–1519.
59 Takabatake N, Nakamura H, Abe S, et al. Circulating leptin in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 1999;159(4 Pt 1):1215–1219.
60 Hansel NN, Gao L, Rafaels NM, et al. Leptin receptor polymorphisms and lung function decline in COPD. *Eur Respir J*. 2009;34(1):103–110.
61 Bruno A, Chanez P, Chiappara G, et al. Does leptin play a cytokine-like role within the airways of COPD patients? *Eur Respir J*. 2005;26(3):398–405.
62 Vernooy JHJ, Drummen NEA, Suylen RJ van, et al. Enhanced pulmonary leptin expression in patients with severe COPD and asymptomatic smokers. *Thorax*. 2009;64(1):26–32.
63 Sueblinvong V, Liangpunsakul S. Relationship between serum leptin and chronic obstructive pulmonary disease in US adults: results from NHANESIII. *J Investig Med*. 2014;62(7):934–937.
64 Curtis JL. Queens beat one-eyed jacks, but nobody's played the ace yet. Adipokines as chronic obstructive pulmonary disease biomarkers. *Ann Am Thorac Soc*. 2015;12(7):971–973.
65 Oh YM, Jeong BH, Woo SY, et al. Association of plasma adipokines with chronic obstructive pulmonary disease severity and progression. *Ann Am Thorac Soc*. 2015;12(7):1005–1012.
66 Stanojevic S, Kaminsky DA, Miller MR, et al. ERS/ATS technical standard on interpretive strategies for routine lung function tests. *Eur Respir J*. 2022;60(1):2101499.
67 Wan ES, DeMeo DL, Hersh CP, et al. Clinical predictors of frequent exacerbations in subjects with severe chronic obstructive pulmonary disease (COPD). *Respir Med*. 2011;105(4):588–594.
68 Wan ES, Fortis S, Regan EA, et al. Longitudinal phenotypes and mortality in preserved ratio impaired spirometry in the COPDGene study. *Am J Respir Crit Care Med*. 2018;198(11):1397–1405.
69 Wan ES, Castaldi PJ, Cho MH, et al. Epidemiology, genetics, and subtyping of preserved ratio impaired spirometry (PRISm) in COPDGene. *Respir Res*. 2014;15:89.
70 Bhatt SP, Balte PP, Schwartz JE, et al. Discriminative accuracy of FEV 1 :FVC thresholds for COPD-related hospitalization and mortality. *JAMA*. 2019;321(24):2438.
71 Moll M, Qiao D, Regan EA, et al. Machine learning and prediction of all-cause mortality in COPD. *Chest*. 2020;158(3):952–964.