Original research

Pulmonary emphysema subtypes defined by unsupervised machine learning on CT scans

Elsa D Angelini,[1,2,3] Jie Yang,[1] Pallavi P Balte,[4] Eric A Hoffman,[5] Ani W Manichaikul ,[6] Yifei Sun,[7] Wei Shen,[8,9] John H M Austin,[10] Norrina B Allen,[11] Eugene R Bleecker,[12] Russell Bowler,[13] Michael H Cho ,[14,15] Christopher S Cooper,[16] David Couper,[17] Mark T Dransfield,[18] Christine Kim Garcia ,[4] MeiLan K Han,[19] Nadia N Hansel,[20] Emlyn Hughes,[21] David R Jacobs,[22] Silva Kasela,[23,24] Joel Daniel Kaufman ,[25] John Shinn Kim ,[4,26] Tuuli Lappalainen,[23] Joao Lima,[20] Daniel Malinsky,[7] Fernando J Martinez,[27] Elizabeth C Oelsner,[4] Victor E Ortega,[28] Robert Paine,[29] Wendy Post,[20] Tess D Pottinger,[4] Martin R Prince,[30] Stephen S Rich,[6] Edwin K Silverman,[14] Benjamin M Smith,[4,31] Andrew J Swift,[4,32] Karol E Watson,[16] Prescott G Woodruff,[33] Andrew F Laine,[1,9,10] R Graham Barr[4,34]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/thorax-2022-219158).

For numbered affiliations see end of article.

Correspondence to
Dr R Graham Barr, Department of Medicine, Columbia University Irving Medical Center, New York, NY 10032, USA; rgb9@columbia.edu

EDA and JY contributed equally. AFL and RGB contributed equally.

Received 3 May 2022
Accepted 3 May 2023
Published Online First 2 June 2023

► http://dx.doi.org/10.1136/thorax-2023-220458

To cite: Angelini ED, Yang J, Balte PP, *et al*. *Thorax* 2023;**78**:1067–1079.

ABSTRACT

Background Treatment and preventative advances for chronic obstructive pulmonary disease (COPD) have been slow due, in part, to limited subphenotypes. We tested if unsupervised machine learning on CT images would discover CT emphysema subtypes with distinct characteristics, prognoses and genetic associations.

Methods New CT emphysema subtypes were identified by unsupervised machine learning on only the texture and location of emphysematous regions on CT scans from 2853 participants in the Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS), a COPD case–control study, followed by data reduction. Subtypes were compared with symptoms and physiology among 2949 participants in the population-based Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study and with prognosis among 6658 MESA participants. Associations with genome-wide single-nucleotide-polymorphisms were examined.

Results The algorithm discovered six reproducible (interlearner intraclass correlation coefficient, 0.91–1.00) CT emphysema subtypes. The most common subtype in SPIROMICS, the combined bronchitis-apical subtype, was associated with chronic bronchitis, accelerated lung function decline, hospitalisations, deaths, incident airflow limitation and a gene variant near *DRD1*, which is implicated in mucin hypersecretion ($p=1.1\times10^{-8}$). The second, the diffuse subtype was associated with lower weight, respiratory hospitalisations and deaths, and incident airflow limitation. The third was associated with age only. The fourth and fifth visually resembled combined pulmonary fibrosis emphysema and had distinct symptoms, physiology, prognosis and genetic associations. The sixth visually resembled vanishing lung syndrome.

Conclusion Large-scale unsupervised machine learning on CT scans defined six reproducible, familiar CT emphysema subtypes that suggest paths to specific diagnosis and personalised therapies in COPD and pre-COPD.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Chronic obstructive pulmonary disease (COPD) and emphysema have long been recognised as heterogeneous, overlapping diseases and some patients with non-obstructive emphysema, or 'pre-COPD,' may progress to COPD; yet modern unsupervised machine learning methods have not been applied at scale to the vast amount of imaging data in contemporary chest CT scans in order to subphenotype emphysema.

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) was the third-leading cause of death globally in 2019.[1] Despite identification of hundreds of genetic loci for COPD,[2] which is defined by chronic airflow limitation,[3] personalised therapies are lacking for most patients due, in part, to a lack of robust subphenotyping.

Historical attempts to subphenotype COPD included pulmonary emphysema, defined by enlargement and destruction of alveoli, and chronic bronchitis, defined by chronic cough and phlegm.[4 5] However, many patients with COPD have neither subphenotype and targeted treatments are limited.

Paradoxically, many individuals who do not have COPD have emphysema or chronic bronchitis,[6–8] which has recently been termed 'pre-COPD.'[3] Emphysema on CT is predictive of morbidity and mortality independent of lung function,[6 9–11] yet it remains uncertain which 'pre-COPD' phenotypes progress to COPD.[3]

Emphysema itself was subdivided into centrilobular, panlobular and paraseptal emphysema based on 142 autopsies[12 13]; yet these subtypes are read with limited reproducibility by radiologists,[14 15] ignored or altered in guidelines,[3 16] and little-used in practice. Hence, traditional subtypes do not provide gold standards and new approaches are warranted.

WHAT THIS STUDY ADDS

⇒ Unsupervised machine learning (clustering) on the texture and anatomical location of millions of emphysematous regions on chest CT scans, followed by data reduction, revealed six CT emphysema subtypes, several of which closely resemble earlier clinical descriptions of COPD subphenotypes. A combined bronchitis-apical emphysema subtype was characterised by symptoms of chronic bronchitis, accelerated lung function decline, increased all-cause mortality and, among those with normal lung function, incident airflow limitation; it was also associated with a gene variant relevant to nicotinic pathways and mucin hypersecretion. A diffuse emphysema subtype was associated with wasting, respiratory hospitalisations and deaths and, among those with normal lung function, incident airflow limitation. An obstructive combined fibrosis pulmonary emphysema subtype was largely asymptomatic but associated with respiratory deaths. The other three CT emphysema subtypes had distinct physiological or genetic associations.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These precise new CT emphysema subtypes have differential prognosis and may suggest paths to more specific diagnosis and personalised therapies in COPD and in pre-COPD.

Unsupervised machine learning is a machine learning approach used to discover naturally occurring clusters without reference to preassigned gold standards.[17] Attempts to subphenotype COPD using unsupervised clustering of symptoms,[18] lung function,[19 20] 'omics[21] and standard CT measures[22 23] have generally not yielded robust, familiar subtypes, possibly due to use of limited variables and samples. Unsupervised machine learning is most powerful when applied at scale to high-dimensional data like research chest CT scans, which provide 20–30 megavoxel, 3-dimensional representations of the entire lung at submillimeter resolution. Lung CT images have not, to our knowledge, been used to learn completely new emphysema subtypes at scale.

We hypothesised that application of a custom-built unsupervised machine learning algorithm[24] to cluster the texture and anatomical location of emphysematous regions on thousands of CT scans, followed by data reduction, would allow robust learning in vivo of new CT emphysema subtypes with distinct characteristics, prognoses and genetic associations. The learning used CT images only so we could examine clinical and genetic associations independent of the learning.

METHODS

The machine learning algorithm was applied to the Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS), which recruited 2783 COPD cases and controls, 40–80 years old with ≥20 pack-years and 200 non-smoking controls in 2010–2015,[25] initially on random 50% subsamples to test reproducibility (figure 1).

Data reduction to CT emphysema subtypes was performed in SPIROMICS and the population-based Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study, which acquired full-lung CT scans in 2010–2012 for 3128 MESA participants.[26] Results were confirmed longitudinally in a subset of 196 MESA Lung participants with repeat CT scans (of 317 oversampled for COPD and with 10+ pack-years[15]).

Primary descriptive analyses of clinical characteristics of CT emphysema subtypes were evaluated in the MESA Lung Study.

Events analyses were performed in MESA, which acquired cardiac CT scans for 6814 whites, blacks, Hispanics and Asians in 2000–2002 with follow-up through 2018; incident airflow limitation was examined among participants without airflow limitation and with repeated spirometry.

Genetic discovery analyses were performed in SPIROMICS, given its greater disease severity. Replication was performed in the MESA SHARe Study, which composed of MESA plus 1595 black and Hispanic family members and 257 other participants with cardiac CT scans,[27] and the Genetic Epidemiology of COPD (COPDGene) Study, which recruited 10 192 non-Hispanic white and black COPD cases and controls ages 40–81 years with ≥10 pack-years.[28]

CT scanning

SPIROMICS and MESA Lung used the same inspiratory high-resolution full-lung CT protocol: 120 kVp, 0.625–0.75 mm slice thickness, 0.5 s rotation time.[29] MESA and MESA SHARe acquired cardiac CT scans, which imaged the lower two-thirds of the lungs.[30] COPDGene performed full-lung CTs following the COPDGene protocol.[28]

Unsupervised machine learning and data reduction

Discover of possible emphysema subtypes

The unsupervised machine learning algorithm was designed to define possible emphysema subtypes, also called spatial lung texture patterns,[24] and was applied blinded to all clinical information including traditional emphysema subtypes. The target number of possible emphysema subtypes was not specified, nor additional direction provided in this step.

In brief, 25×25×25 mm regions of lung were selected for learning if the percentage of emphysema-like lung (voxels <−950 Hounsfield units)[31] in the region was above the upper limit of normal for per cent emphysema, which accounted for variation in body size, demographics, current smoking and scanner manufacturer.[32] Unsupervised learning was performed with two types of image-based features: texture features,[33] using a learnt, dedicated texton codebook to encode patterns of emphysematous regions, and spatial features using lung spatial mapping.[34]

Unsupervised discovery was performed at a regional level in two stages: (1) K-means with a spatial distance metric[24] to group emphysematous regions into a selected large number of clusters and (2) grouping of similar clusters into possible emphysema subtypes via Infomap graph partitioning,[35] which uses minimum description length optimisation criteria to define the number of subtypes. This two-stage approach was more reproducible than single-stage approaches.[36–38]

Data reduction of possible emphysema subtypes to CT emphysema subtypes

To infer if sets of possible emphysema subtypes might represent different severities of a single CT emphysema subtype or distinct subtypes, we used hierarchical clustering and t-distributed stochastic neighbor embedding (t-SNE) projection to examine for further clustering at a participant level (online supplemental file 1).[39]

Descriptive naming of CT emphysema subtypes

Two board-certified chest radiologists assigned descriptive names to the CT emphysema subtypes after reviewing representative

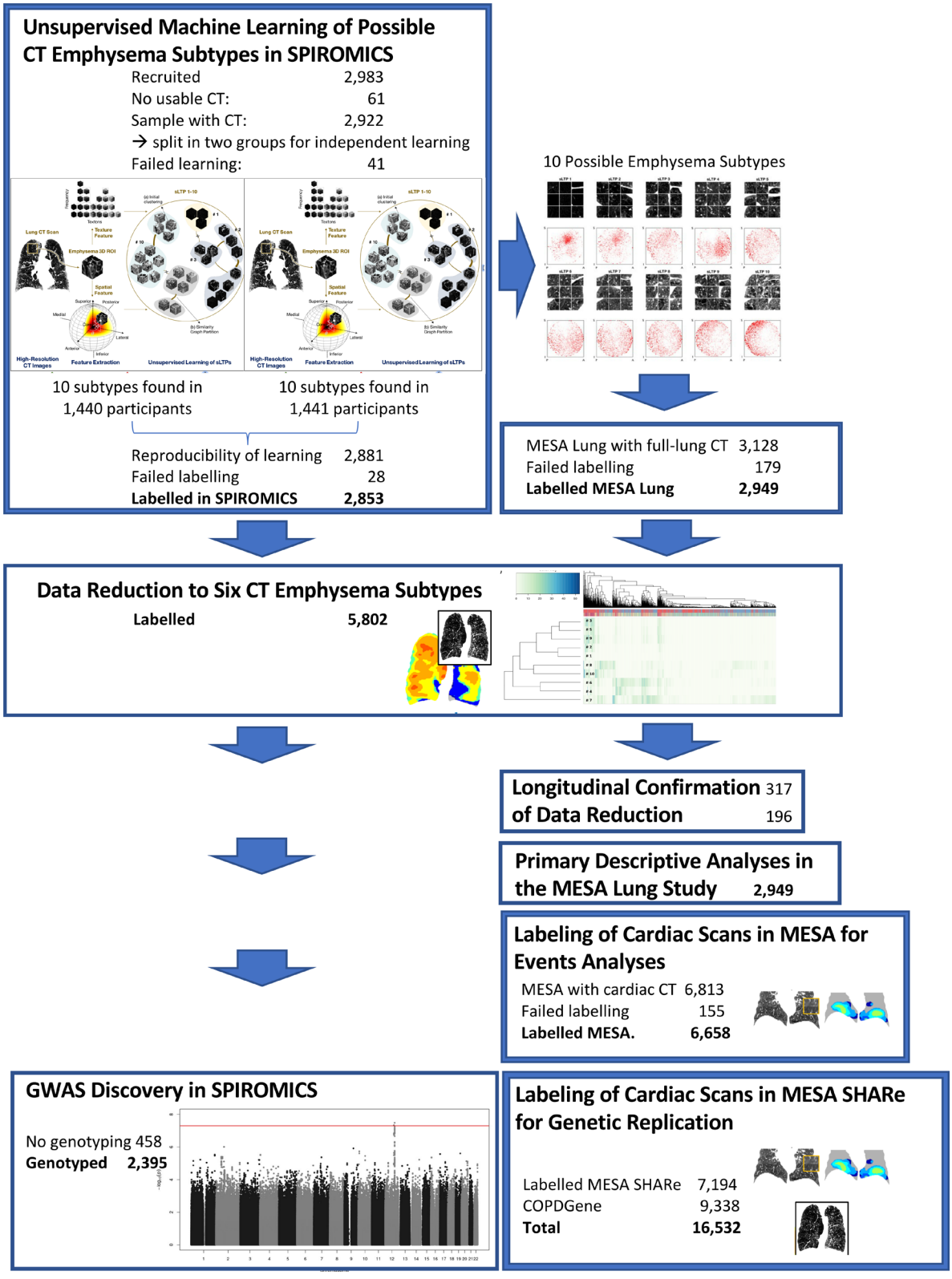

Figure 1 Schema of unsupervised machine learning, data reduction, primary descriptive analyses, events analyses and GWAS. Unsupervised machine learning of possible emphysema subtypes was performed in two independent training sets in SPIROMICS. Both training sets yielded 10 possible emphysema subtypes, and training was repeated on all of SPIROMICS. The resultant 10 possible emphysema subtypes were labelled on MESA Lung CT scans. Data reduction was performed in SPIROMCS and MESA Lung and yielded six CT emphysema subtypes; data reduction was confirmed longitudinally on coregistered CT scans in a subset of the MESA Lung Study oversampled for COPD and smoking. Primary descriptive analyses of these subtypes were performed in the MESA Lung Study. Cardiac scans in MESA were labelled for the Event Analyses in MESA. GWAS Discovery was performed in SPIROMICS; replication of genetic results occurred on labelled cardiac scans in MESA and MESA SHARe and in COPDGene. COPDGene, Genetic Epidemiology of Chronic Obstructive Pulmonary Disease; MESA, Multi-Ethnic Study of Atherosclerosis; SPIROMICS, Subpopulations and Intermediate Outcome Measures in COPD Study. GWAS, genome-wide association study

Table 1 Characteristics of participants in SPIROMICS and the MESA Lung Study

	SPIROMICS (n=2853)	MESA Lung (n=2949)
Age—years	63.0±9.2	69.4±9.3
Male sex—no (%)	1515 (53.1%)	1417 (48.1%)
Race/ethnicity, no (%)		
White	2087 (74.1%)	1123 (38.1%)
Black	550 (19.3%)	803 (27.2%)
Hispanic	148 (5.2%)	631 (21.4%)
Asian	33 (1.2%)	392 (13.3%)
Height—m	1.7±0.1	1.65±0.10
Weight—kg	80.9±18.0	78.1±17.4
BMI—kg/m^2	28.0±5.3	28.4±5.4
Smoking status—no (%)		
Never	198 (6.9%)	1341 (45.8%)
Former	1609 (56.4%)	1371 (46.8%)
Current	1046 (36.7%)	219 (7.5%)
Pack-years, among ever-smokers—median (IQR)	43.0 (31.0–60.0)	14.5 (3.0–33.0)
FEV1, per cent predicted	75.1±26.7	94.9±22.9
FVC, per cent predicted	91.7±18.0	97.2±22.5
FEV1/FVC	0.59±0.16	0.74±0.09
COPD—no (%)	1760 (61.7%)	446 (16.9%)
GOLD 1	380 (21.6%)	239 (53.6%)
GOLD 2	787 (44.8%)	182 (40.8%)
GOLD 3	412 (23.5%)	25 (5.6%)
GOLD 4	178 (10.1%)	0
Total lung volume—mL	5871±1454	4791±1283
Per cent emphysema—%	7.5±10.1	2.5±3.3
Traditional emphysema subtype—no (%)*		
Centrilobular emphysema	804 (93.7)	530 (18.0)
Panlobular emphysema	44 (5.1)	90 (3.1)
Paraseptal emphysema	754 (88.0)	384 (13.0)
Airway wall thickness—mm	1.44±0.42	1.02±0.24
Diasynapsis (CT-assessed airway-to-lung ratio)	0.032±0.004	0.033±0.004
Small airway count (N)	--	30.8±14.9
Interstitial lung abnormalities—no (%)*	252 (25.3)	276 (12.1)
Total pulmonary vascular volume per cent	2.91±0.36	2.70±0.27

*Traditional emphysema subtypes and interstitial lung abnormalities read in SPIROMICS for a subset of 804–857 and 999 participants, respectively.
BMI, body mass index; COPD, chronic obstructive pulmonary disease; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; GOLD, Global Initiative for Chronic Obstructive Lung Disease; MESA, Multi-Ethnic Study of Atherosclerosis; SPIROMICS, Subpopulations and Intermediate Outcome Measures in COPD Study.

examples and anatomic locations (online supplemental figure S1) with consideration of physiological and demographic correlates. Fibrosis, and its colocalisation with emphysema, were determined qualitatively.

Labelling CT emphysema subtypes on cardiac CT scans

We developed a deep learning method using supervised domain adaptation with adversarial learning to label the scanned lung on cardiac CT scans (online supplemental file 1)[40] to increase power for events analyses and genetic replication (figure 1).

Additional measures

Traditional emphysema subtypes and interstitial lung abnormalities (ILAs) were read by board-certified radiologists following standardised protocols.[15 41]

Dyspnoea was assessed using the modified Medical Research Council (mMRC) scale. Chronic bronchitis was defined following MRC criteria.[42]

Spirometry was performed at baseline in SPIROMICS and in 2004–2006, 2010–2012 and 2017–2018 in the MESA Lung Study following American Thoracic Society/European Respiratory Society recommendations.[43] COPD was defined as post-bronchodilator, and airflow limitation as prebronchodilator, forced expiratory volume in 1 s (FEV_1)-to-forced vital capacity (FVC) ratio less than 0.7.[3]

Other lung structure measures of per cent $emphysema_{-950\ HU}$, total lung volume (TLV), airway wall thickness (AWT), small airway count (SAC), dysanapsis, total pulmonary vascular volume (TPVV) and, in SPIROMICS, functional small airways disease were assessed at a single reading centre.[26 29 31 32 44–46]

Exacerbations were self-reported in SPIROMICS. Hospitalisations and deaths from chronic lower respiratory diseases (CLRD) were adjudicated in MESA from 2000 to 2018 with 98% completeness for mortality.[47]

Consenting participants were genotyped with genome-wide arrays (online supplemental file 1). Genome-wide imputation was performed using the Michigan Imputation Server. Colocalisation was performed using expression quantitative trait loci (eQTLs) from the Genotype-Tissue Expression (GTEx).[48]

Statistical analysis

Reproducibility of learning was assessed at a regional level on random test emphysematous regions in SPIROMICS with a regional-level Dice coefficient. Participant-level percentages of each subtype were calculated by summing across lung regions and dividing by TLV, similar to the calculation for per cent emphysema. Participant-level reproducibility of learning was calculated with intraclass correlation coefficients (ICCs).

Generalised linear regression was used to evaluate associations of CT emphysema subtypes at a participant level with demographics, symptoms, lung structure and physiology; mixed linear models were used for lung function decline; and Cox proportional hazards models were used for events. The primary models included demographic, anthropomorphic and smoking potential confounders following a causal framework; CT manufacturer was also included as unmeasured site-level confounders would likely be blocked by this variable (online supplemental figure S1). In a second model, other CT emphysema subtypes that might confound relationships were included. Subsequent analyses adjusted for other lung structure measures and lung function. Analyses were repeated in SPIROMICS with adjustment for recruitment strata given its case–control design.

Genome-wide association analyses were performed with similar adjustment (online supplemental file 1). Primary replication of identified single-nucleotide polymorphism (SNPs) was performed in the race/ethnic group in which they were discovered. Colocalisation of replicated variants and eQTLs used the coloc method (online supplemental file 1).[49 50]

Statistical significance was evaluated with 95% CIs for epidemiological analyses and defined by Bonferroni-corrected $p<5\times10^{-8}$ for genome-wide analyses.

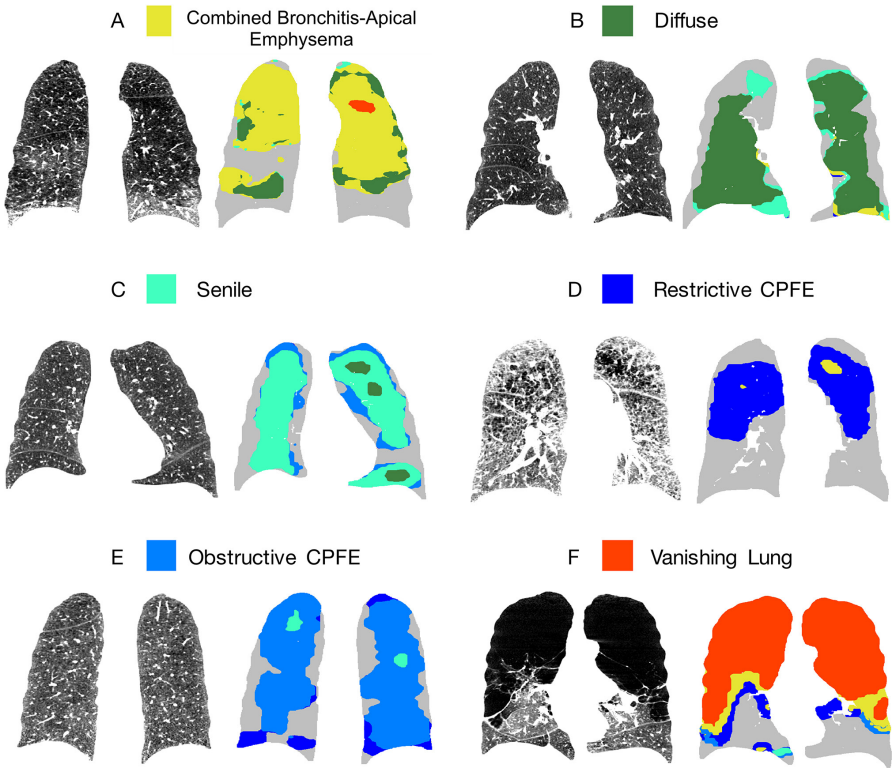

Figure 2 Representative visual illustrations of the six CT emphysema subtypes.Coronal views of lungs on CT scans and the corresponding labelled masks with the discovered CT emphysema subtypes on predominantly affected sample cases (ie, with proportion of a certain CT emphysema subtype being much larger than any other). Colour coding of CT emphysema subtypes is the same across examples; grey labelling denotes non-emphysematous regions. CPFE, combined pulmonary fibrosis/emphysema.

RESULTS

SPIROMICS participants had a median of 43 pack-years, 62% had COPD, 0.4% were PiZZ and the race/ethnic distribution was 74.1% white, 19.3% black, 5.2% Hispanic and 1.2% Asian (table 1). The MESA Lung Study included 54% participants with a smoking history (median 14 pack-years), 16.9% had COPD and the race/ethnic distribution was 38.1% white, 27.2% black, 21.4% Hispanic and 13.3% Asian. MESA and MESA SHARe were similarly multiethnic; COPDGene was biracial (online supplemental table S1).

Unsupervised learning and data reduction to CT emphysema subtypes

SPIROMICS CT scans had an average of 624±350 emphysematous regions per scan covering most of the lung volume for 2922

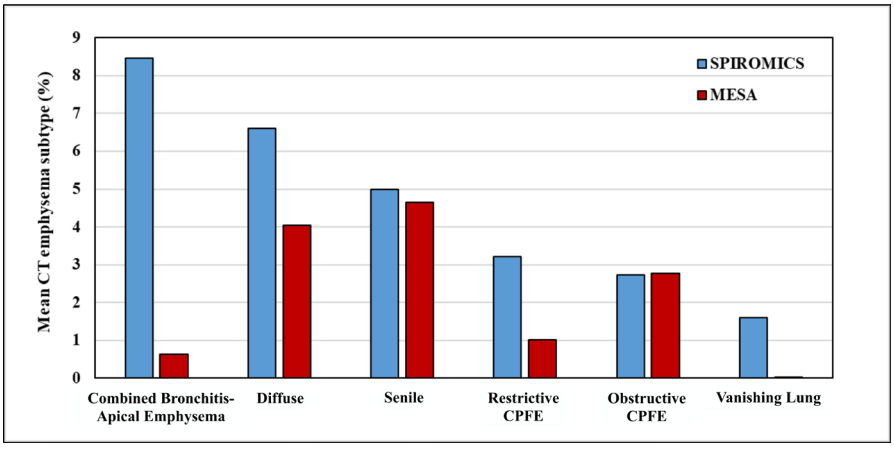

Figure 3 Distributions of the six discovered CT emphysema subtypes in SPIROMICS and the MESA Lung Study. Mean percentages of CT emphysema subtypes in SPIROMICS, a COPD case–control study of 2655 participants with 20 or more pack-years of smoking (median pack-years 43.0; 66.2% with COPD) and 198 non-smoking controls, and in the MESA Lung Study, a population-based study of 2949 participants, 54.2% of whom had ever smoked cigarettes (median pack-years 14.5) and 16.9% with COPD. COPD, chronic obstructive pulmonary disease; CPFE, combined pulmonary fibrosis/emphysema; MESA, Multi-Ethnic Study of Atherosclerosis; SPIROMICS, Subpopulations and Intermediate Outcome Measures in COPD Study.

participants, which yielded over 1.8 million regions for learning. Application of the unsupervised machine learning algorithm to a random 50% of scans yielded 10 possible emphysema subtypes (online supplemental figure S2). Repeating the learning independently on the other 50% also yielded 10 possible emphysema subtypes. Agreement in learning was high: regional-level dice=0.82, participant-level ICC 0.89–1.00 (online supplemental table S2).

Hierarchical clustering suggested that some subtypes overlapped and data dimension reduction suggested that the 10 possible subtypes clustered into 6 CT emphysema subtypes (online supplemental figure S3). For the six CT emphysema subtypes, agreement in learning was also high (ICC 0.91–1.00; online supplemental table S3) and labelling was reproducible (ICC 1.00 for all).

Longitudinal evaluation over 6 years of regions-of-interest on co-registered CT scans confirmed, that, at a regional level, possible emphysema subtypes clustered by t-SNE tended to progress from one to another within the same CT emphysema subtype. In contrast, unclustered possible emphysema subtypes remained distinct or developed from normal lung (online supplemental figure S4).

Qualitative visual description

The resultant six CT emphysema subtypes are illustrated in figure 2. The combined bronchitic-apical emphysema (CBaE) subtype had a predominantly apical distribution with vascular changes. The diffuse subtype had a diffuse distribution with less parenchymal destruction and apical sparing. The senile had homogeneously reduced attenuation. The restrictive combined pulmonary fibrosis/emphysema (CPFE) subtype had distinct and discrete small holes at the level of the secondary pulmonary lobule in predominantly apical and posterior but also inferior regions. The obstructive CPFE subtype had diffuse, patchy emphysema with intermingled regions of fibrosis. The vanishing lung subtype was predominantly apical with bullous emphysema when severe; when less severe, it had prominent lobular septal, reduced parenchyma and few vessels.

Comparison with traditional emphysema subtypes

In the subset of 317 MESA Lung participants oversampled for COPD and smoking, the distribution of CT emphysema subtypes was approximately similar to SPIROMICS (online supplemental table S4). At a participant level, centrilobular emphysema was positively associated and overlapped predominantly with CBaE and restrictive CPFE subtypes. Panlobular emphysema was associated with CBaE and vanishing lung subtypes. Paraseptal emphysema was positively associated with restrictive CPFE and vanishing lung subtypes. The diffuse and obstructive CPFE subtypes were not independently recognised by radiologists.

Clinical and physiological characteristics

The CBaE subtype was much more common in SPIROMICS than in the MESA Lung Study (figure 3). It was associated independently with smoking history (table 2) and symptoms of dyspnoea and—unique among subtypes—chronic bronchitis (OR 1.9 per 10 percentage point increase in CBaE, 95% CI 1.2 to 3.0; figure 4; online supplemental table S5). It was characterised by large cross-sectional decrements in lung function (eg, −309 mL in FEV_1 per 10 percentage point increase in CBaE, 95% CI −389 to −229) but no difference in TLV. In longitudinal analyses, it was associated with decline in FEV_1 (−13.2 mL/year per 10 percentage point, 95% CI −21.7 to −4.8), the FVC and the FEV_1/FVC ratio. These findings were little changed with adjustment for other measures of lung structure and function (online supplemental table S6). The *CBaE* subtype was also associated with a 2–3 fold independent increase in risk of CLRD hospitalisations, CLRD mortality and all-cause mortality and, among participants with normal lung function, incident airflow limitation (table 3). These findings were independent of AWT, ILAs, per cent emphysema and, for CLRD hospitalisations and incident airflow limitation, lung function (online supplemental table S7). In SPIROMICS, it was also associated with worse symptom scores, reduced exercise capacity, desaturation on exertion, increased haemoglobin and exacerbations (online supplemental table S8).

The second most common subtype in SPIROMICS, the diffuse subtype, was also common in MESA Lung (figure 3). Greater age, male sex, white race/ethnicity and lower body mass index (BMI) but not smoking were associated with the diffuse subtype (table 2). It was associated with few symptoms; the FEV_1/FVC ratio was lower cross-sectionally; AWT and TPVV were reduced; the haemoglobin, FVC and TLV were greater (eg, 487 mL per 10 percentage point, 95% CI 448 to 526 mL); and differences in lung function decline were more modest (figure 4; online supplemental table S5). Findings were similar with adjustment for other lung structure measures (online supplemental table S6). The diffuse subtype was associated with an approximately 50% increase in risk for CLRD hospitalisations and CLRD mortality, lower all-cause mortality and, among participants with normal lung function, incident airflow limitation (table 3). These findings were independent of AWT and ILAs but attenuated by per cent emphysema (online supplemental table S7). In SPIROMICS, it was associated with worse symptom scores, hypoxaemia, desaturation on exertion and exacerbations (online supplemental table S8).

The senile subtype was equally common in the two studies (figure 3). Greater age was associated with it (table 2), and it had similar physiological changes to the diffuse subtype but not the poor prognosis (figure 4, table 3).

The restrictive CPFE subtype was more common in SPIROMICS than in the MESA Lung Study (figure 3). It had similar symptomatology to the CBaE subtype but was more common among women and non-whites and was associated with higher BMI, restrictive spirometry, reduced SAC and TLV and greater ILAs (table 2, figure 4). Despite its high symptom burden, it was not independently associated with hospitalisations or mortality (table 3). Findings were similar in SPIROMICS (online supplemental table S8).

The obstructive CPFE subtype was equally common in the two studies (figure 3). Female sex, black and Asian race/ethnicities, and higher BMI were associated independently with it (table 2). It was associated with obstructive spirometry, reduced AWT and greater TLV cross-sectionally (figure 4). In longitudinal analyses, it was associated with significant increases in the FEV_1 and FVC (figure 4, online supplemental table S5) and an 80% increase in risk of CLRD mortality (table 3). The latter finding was independent of AWT and ILAs but attenuated by per cent emphysema (online supplemental table S7).

The vanishing lung subtype occurred mainly in SPIROMICS (figure 3) and was independently associated with dyspnoea, desaturation on exertion and large increases in lung volumes (table 2 and online supplemental table S6.

There were modest differences for some CT emphysema subtypes by CT manufacturer (online supplemental table S9).

Table 2 Associations of demographic factors and smoking history with CT emphysema subtypes in the MESA Lung Study

N=2949	Combined bronchitis-apical emphysema ß (95% CI)	Diffuse emphysema ß (95% CI)	Senile emphysema ß (95% CI)	Restrictive CPFE ß (95% CI)	Obstructive CPFE ß (95% CI)	Vanishing lung emphysema ß (95% CI)
Age, years						
Unadjusted	**0.3 (0.2, 0.4)**	**1.2 (0.9, 1.5)**	**0.5 (0.3, 0.7)**	**0.1 (0.05, 0.24)**	0.05 (–0.1, 0.2)	0.02 (–0.003, 0.03)
Model 1	**0.2 (0.1, 0.3)**	**0.8 (0.5, 1.0)**	**0.4 (0.2, 0.6)**	**0.2 (0.1, 0.3)**	**0.2 (0.1, 0.4)**	0.001 (–0.01, 0.03)
Model 2	0.002 (–0.1, 0.1)	**0.7 (0.4, 1.0)**	**0.3 (0.1, 0.5)**	**0.1 (0.05, 0.2)**	**0.2 (0.02, 0.3)**	–0.001 (–0.01, 0.01)
Sex, male						
Unadjusted	**0.6 (0.3, 0.8)**	**38 (33, 44)**	**7.6 (3.9, 11)**	**–7.0 (–8.7, –5.2)**	**–23 (–25, –20)**	**0.6 (0.2, 0.9)**
Model 1	**3.0 (0.6, 5.5)**	**31.7 (26.5, 36.9)**	3.7 (–0.04, 7.5)	**–8.1 (–9.8, –6.3)**	**–22.9 (–25.6, –20.2)**	**0.4 (0.2, 0.7)**
Model 2	0.03 (–1.6, 1.6)	**21.2 (16.1, 26.4)**	3.6 (–0.3, 7.6)	**–3.4 (–5.0, –1.8)**	**–16.4 (–19.0, –13.7)**	0.2 (–0.03, 0.5)
Race/ethnicity						
Black						
Unadjusted	–0.1 (–3.3, 3.0)	**–35 (–42, –28)**	**–5.9 (–11, –1.2)**	**7.5 (5.3, 9.7)**	**12 (8.9, 16)**	**0.5 (0.04, 0.9)**
Model 1	**3.0 (–0.04, 6.1)**	**–20.5 (–27.0, –13.9)**	–0.8 (–5.5, 3.9)	**5.9 (3.8, 8.1)**	**8.1 (4.8, 11.5)**	**0.7 (0.3, 1.1)**
Model 2	–0.2 (–2.1, 1.8)	**–16.3 (–22.5, –10.2)**	0.2 (–4.4, 4.9)	**3.4 (1.4, 5.3)**	**3.8 (0.6, 6.9)**	**0.3 (0.004, 0.6)**
Hispanic						
Unadjusted	**–5.9 (–9.3, –2.6)**	**–39 (–47, –32)**	–12 (–17, –7)	**5.0 (2.7, 7.4)**	**13 (9.2, 17)**	–0.27 (–0.72, 0.18)
Model 1	–1.3 (–4.6, 2.0)	**–27.1 (–34.0,–20.1)**	**–6.6 (–12, –1.6)**	**5.5 (3.2, 7.8)**	**11.5 (7.9, 15.1)**	0.1 (–0.4, 0.6)
Model 2	–0.2 (–2.3, 1.8)	**–20.1 (–26.7,–13.5)**	**–5.5 (–10.5, –0.5)**	**2.8 (0.7, 4.9)**	**7.0 (3.6, 10.4)**	0.03 (–0.3, 0.3)
Asian						
Unadjusted	–1.2 (–5.2, 2.7)	–6.1 (–15, 2.7)	3.8 (–2.1, 9.8)	1.6 (–1.2, 4.4)	**9.5 (5.0, 14)**	0.1 (–0.5, 0.1)
Model 1	–0.9 (–5.0, 3.2)	**–24.2 (–32.9, –15.5)**	0.3 (–6.0, 6.6)	**8.6 (5.7, 11.4)**	**20.7 (16.2, 25.2)**	0.04 (–0.5, 0.6)
Model 2	–1.0 (–3.5, 1.6)	**–14.9 (–23.1, –6.6)**	0.4 (–5.8, 6.7)	**4.2 (1.6, 6.8)**	**14.4 (10.2, 18.6)**	0.04 (–0.3, 0.4)
Body mass index, kg/m^2						
Unadjusted	**–0.6 (–0.9, –0.4)**	**–3.7 (–4.2, –3.2)**	**–1.3 (–1.7, –1.0)**	**0.6 (0.5, 0.8)**	**1.5 (1.3, 1.8)**	**–0.04 (–0.1, –0.01)**
Model 1	**–0.6 (–0.9, –0.4)**	**–3.1 (–3.6, –2.6)**	**–1.1 (–1.5, –0.8)**	**0.6 (0.4, 0.8)**	**1.5 (1.3, 1.8)**	**–0.05 (–0.1, –0.02)**
Model 2	**–0.2 (–0.3, –0.03)**	**–2.0 (–2.5, –1.5)**	**–0.9 (–1.3, –0.6)**	**0.3 (0.2, 0.5)**	**1.0 (0.8, 1.3)**	–0.01 (–0.03, 0.02)
Smoking, pack-years						
Unadjusted	**0.4 (0.3, 0.4)**	**0.4 (0.3, 0.5)**	**0.3 (0.2, 0.3)**	**0.2 (0.17, 0.25)**	**0.1 (0.1, 0.02)**	**0.03 (0.02, 0.04)**
Model 1	**0.4 (0.3, 0.4)**	**0.2 (0.1, 0.4)**	**0.2 (0.1, 0.3)**	**0.2 (0.2, 0.3)**	**0.2 (0.1, 0.3)**	**0.03 (0.02, 0.04)**
Model 2	**0.1 (0.1, 0.2)**	0.1 (–0.03, 0.2)	0.1 (–0.01, 0.2)	**0.1 (0.06, 0.14)**	**0.1 (0.05, 0.2)**	–0.01 (–0.01, 0.001)

Results were obtained from linear regression models with the CT emphysema subtype as the dependent variable. ß estimates show the difference in CT emphysema subtype per unit of the demographic factors and packyears (1/0 for categorical variables). Statistically significant results are in bold.
Model 1 was adjusted for the variables in the table plus smoking status and scanner manufacturer.
Model 2 additionally adjusted for other CT emphysema subtypes.
CPFE, combined pulmonary fibrosis/emphysema; MESA, Multi-Ethnic Study of Atherosclerosis.

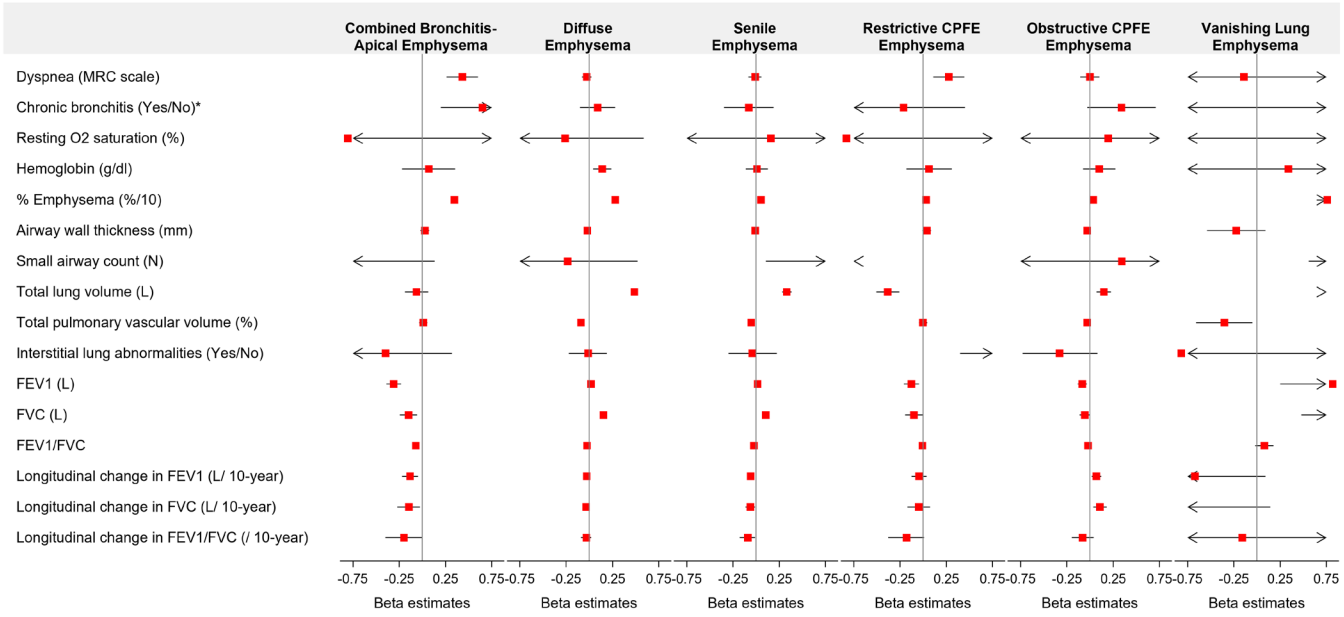

Figure 4 Multivariable associations of CT emphysema subtypes with symptoms, physiology, lung structure and lung function decline in the MESA Lung Study. *β estimates for continuous outcomes show the effect size per 10% increment in CT emphysema subtype, except for per cent emphysema, which is per 1% increment in CT emphysema subtype. The β estimates for chronic bronchitis and interstitial lung abnormalities are the log(ORs). All results adjusted for age, sex, race/ethnicity, height, weight, smoking status, pack-years, scanner manufacturer and other CT emphysema subtypes. CPFE, combined pulmonary fibrosis/emphysema; FEV_1, forced expiratory volume in 1 s; FVC, forced expiratory volume in 1 s; MESA, Multi-Ethnic Study of Atherosclerosis; MRC, Medical Research Council.

Genetic associations

In SPIROMICS, no SNP reached genome-wide significance for the *CBaE* subtype; however, rs35563062 was significantly associated with the lowest attenuation (most severe) of the three possible subtypes that comprise the *CBaE* subtype in White and all participants ($p=1.1\times10^{-8}$; figure 5 and online supplemental table S10). Meta-analysis of replication results was statistically significant in white and all participants (online supplemental table S11). It did not show evidence for colocalisation. The closest gene, *DRD1*, encodes for the dopamine receptor$_1$ (DRD1).

There were no replicated genome-wide significant associations for the diffuse or senile subtypes.

The SNP most significantly associated with the restrictive CPFE subtype in White and all participants (rs113562654, $p=4.5\times10^{-8}$) lies in *NR2C1* (figure 5 and online supplemental table S10). Meta-analysis of replication results was statistically significant in white and all participants (online supplemental table S11) and it colocalised with eQTL for *NR2C1* in GTEx lung tissue (online supplemental figure S5).

Two loci were identified for obstructive CPFE subtype (figure 5 and online supplemental table S10), of which one (rs149784669, $p=4.6\times10^{-9}$), near to *EXOSC*, was unique to and replicated among blacks (online supplemental table S11).

The PI Z variant in *SERPINA1* was not significantly associated with a CT emphysema subtype.

DISCUSSION

Unsupervised machine learning on over 1.8 million emphysematous regions on CT scans defined 6 reproducible CT emphysema subtypes with distinct symptoms, physiology, prognosis and, for 3, replicated genetic associations. The two most common subtypes predicted incident airflow limitation among participants without COPD, improving the specificity of 'pre-COPD.' All resembled early COPD subtypes, which are ignored in contemporary guidelines, and provide precise CT-defined subtypes, some of which suggest avenues to personalised medicine.

The most common emphysema subtype in SPIROMICS was the CBaE subtype, which was read by radiologists as centrilobular or panlobular emphysema. The *CBaE* subtype was strongly related to smoking and uniquely associated with bronchitic symptoms, unchanged TLV and increased haemoglobin—similar to the original description of bronchitic, type B ('blue bloaters') COPD: patients who 'produced large quantities of sputum, … had relatively smaller total lung capacities' and polycythaemia.[5] The *CBaE* subtype also was associated independently with accelerated lung function decline, incident airflow limitation, exacerbations, hospitalisations and all-cause mortality. The original type B subtype applied to few patients; the machine-learnt *CBaE* subtype appears to be a major subset of smoking-related COPD and 'pre-COPD.'

The *CBaE* subtype was associated with a gene variant near *DRD1*. DRD1 is relevant to smoking-related disease as nicotine has dopaminergic effects.[51] DRD1 is present on the airway epithelium, where it increases mucin production and specifically MUC5AC,[52] consistent with the observed bronchitic symptoms with this subtype. MUC5AC is hypothesised to contribute to COPD[53] by causing small airway loss[54] and lung function decline,[55] as observed for this subtype. Dozens of approved drugs target DRD1, suggesting paths towards personalised treatments for the *CBaE* subtype.

The diffuse subtype was associated with few symptoms, lower BMI and higher TLV, similar to the original description of emphysematous, type A ('pink puffers') COPD who had 'little sputum, and rarely showed hypercapnia or recurrent heart failure; their total lung capacities tended to be increased.'[5 56] It also was associated with incident airflow limitation and CLRD hospitalisations and deaths. The diffuse subtype was not recognised

Table 3 Associations of CT emphysema subtypes labelled on cardiac CT scans from 2000 to 2002 with incident clinical events and incident airflow limitation in MESA

			Combined bronchitis-apical emphysema	Diffuse emphysema	Senile Emphysema	Restrictive CPFE	Obstructive CPFE	Vanishing Lung Emphysema
	N Events Person-years	Rate per 10 000 person-years	HR per 10% increment (95% CI)	HR per 10% increment (95% CI)	HR per 10% increment (95% CI)	HR per 10% increment (95% CI)	HR per 10% increment (95% CI)	HR per 10% increment (95% CI)
CLRD hospitalisation (N=6658)								
Unadjusted	148 77 466	19.1	**3.2 (2.8, 3.7)**	**1.6 (1.3, 1.9)**	1.4 (0.9, 2.0)	**1.7 (1.3, 2.2)**	**1.6 (1.2, 2.2)**	**10.7 (5.6, 20.8)**
Model 1			**3.3 (2.8, 3.9)**	**1.9 (1.6, 2.3)**	1.3 (0.9, 1.9)	1.3 (0.9, 1.8)	1.2 (0.9, 1.8)	**10.8 (5.3, 21.9)**
Model 2			**2.9 (2.3, 3.6)**	**1.5 (1.1, 1.9)**	0.8 (0.5, 1.3)	1.0 (0.6, 1.7)	1.4 (0.8, 2.2)	1.1 (0.4, 3.1)
CLRD mortality (N=6658)								
Unadjusted	74	9.5	**3.1 (2.6, 3.6)**	**1.9 (1.6, 2.4)**	**1.9 (1.2, 3.1)**	**1.8 (1.3, 2.6)**	**1.9 (1.4, 2.7)**	**13.0 (6.5, 26.2)**
Model 1	78 096		**2.5 (2.1, 3.0)**	**1.8 (1.5, 2.3)**	1.6 (0.9, 2.7)	**1.6 (1.0, 2.3)**	**1.6 (1.1, 2.2)**	**7.0 (3.2, 15.5)**
Model 2			**2.2 (1.7, 2.9)**	**1.5 (1.1, 2.0)**	0.9 (0.5, 1.6)	0.99 (0.5, 1.9)	**1.8 (1.0, 3.1)**	1.3 (0.4, 4.0)
All-cause mortality (N=6658)								
Unadjusted	1131	144.8	**1.7 (1.5, 1.9)**	1.1 (0.9, 1.2)	1.0 (0.9, 1.2)	1.2 (0.9, 1.4)	1.0 (0.8, 1.3)	**4.3 (2.7, 7.1)**
Model 1	78 096		**1.5 (1.3, 1.7)**	1.0 (0.9, 1.1)	0.9 (0.8, 1.1)	1.0 (0.9, 1.3)	0.9 (0.8, 1.1)	**3.3 (2.0, 5.5)**
Model 2			**1.6 (1.3, 1.8)**	**0.9 (0.8, 0.9)**	1.0 (0.8, 1.2)	1.1 (0.8, 1.4)	0.9 (0.7, 1.2)	1.0 (0.5, 2.1)
Incident airflow limitation (N=2324)								
Unadjusted	364 31 931	114.0	**6.9 (3.0, 15.8)**	**1.5 (1.3, 1.8)**	**1.7 (1.3, 2.1)**	1.3 (0.9, 1.9)	1.2 (0.9, 1.6)	--
Model 1			**7.1 (3.2, 15.7)**	**1.5 (1.3, 1.8)**	**1.6 (1.3, 2.0)**	1.2 (0.8, 1.7)	**1.4 (1.1, 1.8)**	--
Model 2			**2.9 (1.01, 8.5)**	**1.4 (1.1, 1.7)**	1.2 (0.9, 1.6)	1.1 (0.7, 1.7)	1.1 (0.8, 1.6)	--

Airflow limitation defined by prebronchodilator ratio of the forced expiratory volume in one second to the forced vital capacity <0.70. Airflow limitation model for the vanishing lung emphysema subtype did not converge. HRs are per 10% increment in CT emphysema subtype. Statistically significant results are in bold.
Model 1. Adjusted for age, sex, race/ethnicity, height, weight, smoking status, pack-years and scanner manufacturer.
Model 2. Additionally adjusted for other CT emphysema subtypes.
CLRD, Chronic lower respiratory disease; CPFE, combined pulmonary fibrosis/emphysema; MESA, Multi-Ethnic Study of Atherosclerosis.

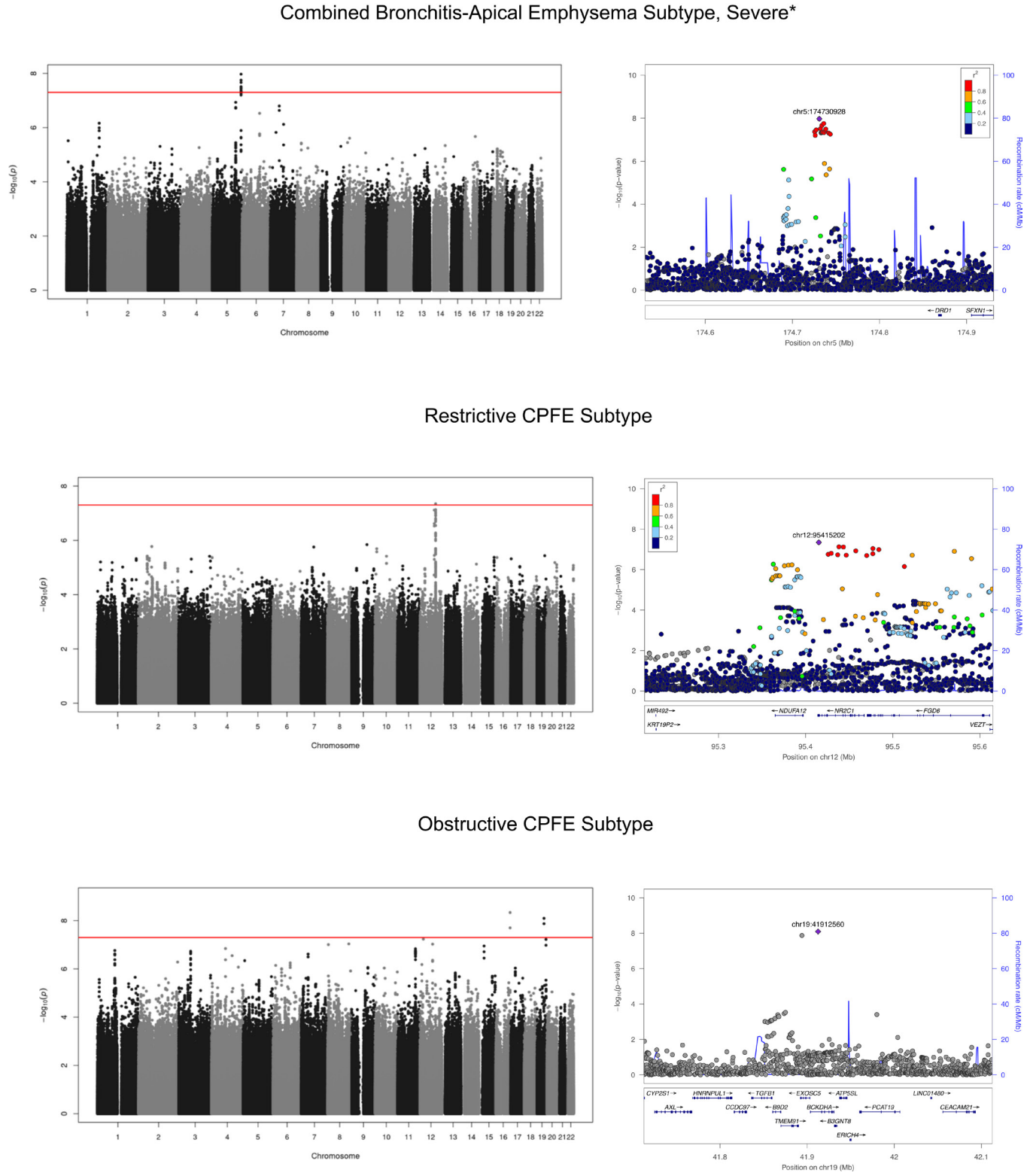

Figure 5 Manhattan and local association plots for the three genome-wide significant, replicated gene variants for three CT emphysema subtypes in SPIROMICS.The red lines show the level of statistical significance ($p=5\times10^{-8}$). The genome-wide significant SNP for the Combined Bronchitis-Apical Emphysema subtype replicated among Whites (p=0.01) and the entire replication sample (p=0.04). The genome-wide significant SNP for the restrictive CPFE subtype replicated among Whites (p=0.01) and the entire replication sample (p=0.04). The first genome-wide significant SNP for the obstructive CPFE subtype on chromosome 19 had variance only among black participants and replicated in this sample (p=0.046). The second genome-wide significant SNP for the obstructive CPFE subtype on chromosome 16 did not replicate. There were no significant replicated genetic associations for the diffuse and senile CT emphysema subtypes (not shown). *Results are shown for the lowest attenuation (most severe) of the three preliminary subtypes that comprise the Combined Bronchitis-Apical Emphysema subtype. COPD, combined pulmonary fibrosis/emphysema; CPFE, combined pulmonary fibrosis/emphysema; SNP, single-nucleotide polymorphism; SPIROMICS, Subpopulations and Intermediate Outcome Measures in COPD Study.

independently by radiologists but was strongly correlated with per cent emphysema (r=0.88). This homogeneous loss of lung tissue may relate to microvascular disease[57 58] or environmental exposures.[59] The original type A subtype has largely disappeared from the literature; the machine-learnt diffuse subtype appears to be a major subset of COPD and 'pre-COPD' unrelated to smoking.

The senile subtype was age-related but not associated independently with morbidity or mortality. The concept of a benign, age-related emphysema is long-standing in the literature[60–63] but, to our knowledge, has not been specifically defined previously.

Two CPFE subtypes were more common among non-white participants: one common in participants with a smoking history and associated with restrictive physiology; the other common in the general population and associated with obstructive physiology. The first was classified by radiologists as centrilobular or paraseptal emphysema; the second was not recognised independently. Distinct gene variants were identified for each. The first is in *NR2C1*, which is close to *FGD6*, which is implicated in macular degeneration, another smoking-related disease.[64] The other, which was observed only in Black participants, is near *EXOSC5*, which is expressed in the lung[65] and codes for exosome component 5, which is implicated in lung diseases.[66 67] CPFE tends to have high symptom burden,[68] consistent with our findings, and restrictive physiology is relevant in COPD.[3 69]

The last, rare CT emphysema subtype occurred only with a smoking history, was bullous, and visually resembled vanishing lung syndrome (giant bullous emphysema).[70]

This is the first report of which we are aware to use large-scale unsupervised learning on CT images to define new CT emphysema subtypes. Our preliminary report[24] yielded similar possible emphysema subtypes but was based on 1/10 the sample size. Unsupervised approaches using an autoencoder[71] and existing CT measures[72] on small subsets of SPIROMICS and a preliminary report using standard texture features in a generative model[73] did not result in familiar subtypes.

Strengths of the current report include automated learning of emphysema subtypes on lung images, high reproducibility of learning, CT emphysema subtypes that echo the older literature, biologically relevant genetic associations and multiethnic discovery and replication.

Nonetheless, data reduction strategies were not as robust as unsupervised machine learning and some CT emphysema subtypes might represent a more severe form of another, although genetic and longitudinal results support the current classification. The distribution of some CT emphysema subtypes varied between SPIROMICS and MESA Lung, which was expected given study design differences. We did not validate the subtypes against histology, preventing cellular-level insights. No gold standard was available, but the mirroring of the classic literature suggests construct validity. Learning was based on cross-sectionally acquired scans, although longitudinal analyses suggested subtypes were relatively stable. CT emphysema subtypes are continuous measures; further work is needed to define thresholds to categorise individuals. Differentiation of CPFE subtypes from traction bronchiectasis and honeycombing was not explicit; however, the predominantly upper lobe and generalised anatomic distributions of the two CPFE subtypes were not typical of them. Events analyses used cardiac CT scans, which may underestimate risk for some subtypes. Some epidemiological associations varied by study, but many were consistent with the classic literature. Not all genetic results colocalised; nonetheless, replicated loci and nearby candidate genes were biologically plausible.

In summary, large-scale unsupervised machine learning applied to lung CT scans defined six novel, reproducible CT emphysema subtypes that bore similarities to previously described but largely discarded subtypes. The CBaE and diffuse subtypes were associated with incident airflow limitation among individuals without COPD and poor outcomes among those with COPD. Additional studies are warranted to test if implicated genes are causal and drugs targeting identified pathways yield personalised strategies for 'pre-COPD' and COPD.

Author affiliations
[1]Department of Biomedical Engineering, Columbia University, New York, New York, USA
[2]LTCI, Institut Polytechnique de Paris, Telecom Paris, Palaiseau, France
[3]NIHR Imperial Biomedical Research Centre, ITMAT Data Science Group, Imperial College, London, UK
[4]Department of Medicine, Columbia University Irving Medical Center, New York, New York, USA
[5]Departments of Radiology, Medicine and Biomedical Engineering, University of Iowa, Iowa City, Iowa, USA
[6]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA
[7]Department of Biostatistics, Columbia University Irving Medical Center, New York, New York, USA
[8]Department of Pediatrics, Institute of Human Nutrition, Columbia University Irving Medical Center, New York, New York, USA
[9]Columbia Magnetic Resonance Research Center (CMRRC), Columbia University Irving Medical Center, New York, New York, USA
[10]Department of Radiology, Columbia University Irving Medical Center, New York, New York, USA
[11]Institute for Public Health and Medicine (IPHAM) - Center for Epidemiology and Population Health, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
[12]Department of Medicine, University of Arizona Health Sciences, Tucson, Arizona, USA
[13]Department of Medicine, National Jewish Health, Denver, Colorado, USA
[14]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[15]Harvard Medical School, Boston, Massachusetts, USA
[16]Department of Medicine, University of California, Los Angeles, California, USA
[17]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, USA
[18]Lung Health Center, University of Alabama, Birmingham, Alabama, USA
[19]Department of Medicine, University of Michigan, Ann Arbor, Michigan, USA
[20]Department of Medicine, Johns Hopkins University, Baltimore, Maryland, USA
[21]Department of Physics, Columbia University, New York, New York, USA
[22]Division of Epidemiology and Community Public Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA
[23]Department of Systems Biology, Columbia University Irving Medical Center, New York, New York, USA
[24]New York Genome Center, New York, New York, USA
[25]Departments of Environmental & Occupational Health Sciences, Medicine, and Epidemiology, University of Washington, Seattle, Washington, USA
[26]Department of Medicine, University of Virginia School of Medicine, Charlottesville, Virginia, USA
[27]Department of Medicine, Cornell University Joan and Sanford I Weill Medical College, New York, New York, USA
[28]Department of Pulmonary Medicine, Mayo Clinic, Phoenix, Arizona, USA
[29]Department of Medicine, University of Utah, Salt Lake City, Utah, USA
[30]Department of Radiology, Cornell University Joan and Sanford I Weill Medical College, New York, New York, USA
[31]Department of Medicine, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
[32]Department of Infection, Immunity and Cardiovascular Disease, The University of Sheffield, Sheffield, UK
[33]Department of Medicine, University of California, San Francisco, California, USA
[34]Department of Epidemiology, Columbia University Irving Medical Center, New York, New York, USA

Twitter John Shinn Kim @jskim8223

Acknowledgements The authors thank the other investigators, the staff and the participants of the SPIROMICS, MESA Lung Study and COPD Gene Study for their valuable contributions. A full list of participating MESA investigators and institutions and how to access MESA data can be found at www.mesa-nhlbi.org. More information about SPIROMICS and how to access SPIROMICS data is at www.

spiromics.org. We would like to acknowledge the following current and former investigators of the SPIROMICS sites and reading centres: Neil E Alexis, PhD; Wayne NES Anderson, PhD; R Graham Barr, MD, DrPH; Eugene R Bleecker, MD; Richard C Boucher, MD; Russell P Bowler, MD, PhD; Elizabeth E Carretta, MPH; Stephanie A Christenson, MD; Alejandro P Comellas, MD; Christopher B Cooper, MD, PhD; David J Couper, PhD; Gerard J Criner, MD; Ronald G Crystal, MD; Jeffrey L Curtis, MD; Claire M Doerschuk, MD; Mark T Dransfield, MD; Christine M Freeman, PhD; MeiLan K Han, MD, MS; Nadia N Hansel, MD, MPH; Annette T Hastie, PhD; Eric A Hoffman, PhD; Robert J Kaner, MD; Richard E Kanner, MD; Eric C Kleerup, MD; Jerry A Krishnan, MD, PhD; Lisa M LaVange, PhD; Stephen C Lazarus, MD; Fernando J Martinez, MD, MS; Deborah A Meyers, PhD; John D Newell Jr, MD; Elizabeth C Oelsner, MD, MPH; Wanda K O'Neal, PhD; Robert Paine, III, MD; Nirupama Putcha, MD, MHS; Stephen I. Rennard, MD; Donald P Tashkin, MD; Mary Beth Scholand, MD; J Michael Wells, MD; Robert A Wise, MD; and Prescott G Woodruff, MD, MPH. The project officers from the Lung Division of the National Heart, Lung, and Blood Institute were Lisa Postow, PhD, Thomas Croxton, PhD, MD and Antonello Punturieri, MD, PhD.

Contributors EA, JY, WS, AL and RGB contributed to the machine learning; PB, YS, DC, JSK, DM and RGB contributed to the epidemiological analyses; EAH, NA, CSC, MTD, CKG, EH, MKH, NNH, DRJ, JDK, JL, FJM, EO, RP, MRP, WP, BS, KEW, PGW and RGB contributed to data collection or funding; AWM, ERB, RB, MC, SK, TL, VEO, TP, SSR and EKS contributed to the genomic analyses; JHMA and AJS provided radiologist interpretations; EA and JY drafted the manuscript; all authors contributed to revisions and provided final approval.

Funding This work was supported by NIH/NHLBI R01-HL121270, R01-HL077612, R01-HL093081, R01-HL142028, R01-HL130506, R01-HL131565, R01-HL103676 and T32-HL144442. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159-69, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881 and DK063491. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. SPIROMICS was supported by contracts from NIH/NHLBI (HHSN268200900013C-20C), which were supplemented by contributions made through the Foundation for the NIH and COPD Foundation from AstraZeneca; Bellerophon Pharmaceuticals; Boehringer-Ingelheim Pharmaceuticals; Chiesi Farmaceutici SpA; Forest Research Institute; GSK; Grifols Therapeutics; Ikaria Nycomed; Takeda Pharmaceutical Company; Novartis Pharmaceuticals Corporation; Regeneron Pharmaceuticals and Sanofi. The COPD Gene Study was supported by NIH grants K12HL120004, R01HL113264, U01HL089856 and P01HL105339. The COPD Gene Study is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.

Competing interests EDA, PPB, AM, YS, WS, JHMA, MHC, DC, EH, DRJ, SK, JDK, TL, JL, ECO, WP, MRP, SSR, EKS, KEW and AFL reports receiving grants from the National Institutes of Health (NIH). JY performed the work at Columbia University but is now an employee of Google. EAH reports receiving grants from the NIH; being a founder and shareholder of VIDA Diagnostics; and holding patents for an apparatus for analysing CT images to determine the presence of pulmonary tissue pathology, an apparatus for image display and analysis, and a method for multiscale meshing of branching biological structures. EBA reports receiving grants from the American Heart Association and the NIH. CBC reports receiving personal fees from GlaxoSmithKline. MTD reports receiving a grant from the NHLBI and personal fees from AstraZeneca, GlaxoSmithKline, Pulmonx, PneumRx/BTG and Quark. MKH reports consulting for GlaxoSmithKline, AstraZeneca and Boehringer Ingelheim receiving research support from Novartis and Sunovion. NNH reports receiving grants from the NIH, Boehringer Ingelheim, and the COPD Foundation. JDK reports receiving grants from US Environmental Protection Agency and the NIH. FJM reports serving on COPD advisory boards for AstraZeneca, Boehringer Ingelheim, Chiesi, GlaxoSmithKline, Sunovion and Teva; serving as a consultant for ProterixBio and Verona; serving on the steering committees of studies sponsored by the NHLBI, AstraZeneca, and GlaxoSmithKline; having served on data safety and monitoring boards of COPD studies supported by Genentech and GlaxoSmithKline. BMS reports receiving grants from the NIH, Canadian Institutes of Health Research (CIHR), Fonds de la recherche en santé du Québec (FRQS), the Research Institute of the McGill University Health Centre, the Quebec Lung Association and AstraZeneca. PGW reports receiving personal fees for consultancy from Theravance, AstraZeneca, Regeneron, Sanofi, Genentech, Roche and Janssen. RGB reports receiving grants from the COPD Foundation, the US Environmental Protection Agency (EPA), the American Lung Association and the NIH.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and this work was approved by the institutional review board of Columbia University Medical Center (AAAE97603). Written informed consent was obtained from all participants. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement SPIROMICS and MESA data are available to the scientific community as described in the Acknowledgements section and on the study websites.

ORCID iDs
Ani W Manichaikul http://orcid.org/0000-0002-5998-795X
Michael H Cho http://orcid.org/0000-0002-4907-1657
Christine Kim Garcia http://orcid.org/0000-0002-0771-1249
Joel Daniel Kaufman http://orcid.org/0000-0003-4174-9037
John Shinn Kim http://orcid.org/0000-0002-8887-150X

REFERENCES

1 World Health Organization. The top 10 causes of death, 2019. Geneva, Switzerland WHO; 2020. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [Accessed 07 Jul 2021].
2 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;51:481–93.
3 Global stategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease - 2023 report: Global initiative for chronic obstructive lung disease; 2022.
4 Baldwin ED, Cournand A, Richards DW. Pulmonary insufficiency; a study of 122 cases of chronic pulmonary emphysema. *Medicine (Baltimore)* 1949;28:201–37.
5 Burrows B, Fletcher CM, Heard BE, *et al*. The emphysematous and bronchial types of chronic airways obstruction. A Clinicopathological study of patients in London and Chicago. *Lancet* 1966;1:830–5.
6 Oelsner EC, Hoffman EA, Folsom AR, *et al*. Association between emphysema-like lung on cardiac computed tomography and mortality in persons without airflow obstruction: a cohort study. *Ann Intern Med* 2014;161:863–73.
7 Woodruff PG, Couper D, Han MK. Symptoms in smokers with preserved pulmonary function. *N Engl J Med* 2016;375:896–7.
8 Balte PP, Chaves PHM, Couper DJ, *et al*. Association of nonobstructive chronic bronchitis with respiratory health outcomes in adults. *JAMA Intern Med* 2020;180:676–86.
9 McAllister DA, Ahmed FS, Austin JHM, *et al*. Emphysema predicts hospitalisation and incident airflow obstruction among older smokers: a prospective cohort study. *PLoS ONE* 2014;9:e93221.
10 Oelsner EC, Carr JJ, Enright PL, *et al*. Per cent emphysema is associated with respiratory and lung cancer mortality in the general population: a cohort study. *Thorax* 2016;71:624–32.
11 Ash SY, San José Estépar R, Fain SB, *et al*. Relationship between emphysema progression at CT and mortality in ever-smokers: results from the Copdgene and ECLIPSE cohorts. *Radiology* 2021;299:222–31.
12 Leopold JG, Gough J. The centrilobular form of hypertrophic emphysema and its relation to chronic Bronchitis. *Thorax* 1957;12:219–35.
13 Edge J, Simon G, Reid L. Peri-Acinar (Paraseptal) emphysema: Its clinical, radiological, and physiological features. *Br J Dis Chest* 1966;60:10–8.
14 Barr RG, Berkowitz EA, Bigazzi F, *et al*. A combined pulmonary-radiology workshop for visual evaluation of COPD: study design, chest CT findings and concordance with quantitative evaluation. *COPD* 2012;9:151–9.
15 Smith BM, Austin JHM, Newell JD, *et al*. Pulmonary emphysema subtypes on computed tomography. The MESA COPD study. *Am J Med* 2014;127:94.
16 Lynch DA, Austin JHM, Hogg JC, *et al*. CT-definable subtypes of chronic obstructive pulmonary disease: a statement of the Fleischner society. *Radiology* 2015;277:192–205.
17 Hinton G, Sejnowski TJ. Unsupervised learning. In: *Foundations of Neural Computation*. MIT Press, 1999.
18 Castaldi PJ, Boueiz A, Yun J, *et al*. Machine learning characterization of COPD subtypes: Insights from the Copdgene study. *Chest* 2020;157:1147–57.
19 Delgado-Eckert E, James A, Meier-Girard D, *et al*. Lung function fluctuation patterns unveil asthma and COPD phenotypes unrelated to type 2 inflammation. *J Allergy Clin Immunol* 2021;148:407–19.
20 Augustin IML, Spruit MA, Houben-Wilke S, *et al*. The respiratory Physiome: clustering based on a comprehensive lung function assessment in patients with COPD. *PLoS ONE* 2018;13:e0201593.
21 Gillenwater LA, Helmi S, Stene E, *et al*. Multi-Omics subtyping pipeline for chronic obstructive pulmonary disease. *PLoS One* 2021;16:e0255337.

22 Zou C, Li F, Choi J, *et al*. Longitudinal imaging-based clusters in former smokers of the copd cohort associate with clinical characteristics: the subpopulations and intermediate outcome measures in copd study (SPIROMICS). *Int J Chron Obstruct Pulmon Dis* 2021;16:1477–96.
23 Young AL, Bragman FJS, Rangelov B, *et al*. Disease progression modeling in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2020;201:294–302.
24 Yang J, Angelini ED, Balte PP, *et al*. Novel subtypes of pulmonary emphysema based on spatially-informed lung texture learning: the multi-ethnic study of Atherosclerosis (MESA) COPD study. *IEEE Trans Med Imaging* 2021;40:3652–62.
25 Couper D, LaVange LM, Han M, *et al*. Design of the subpopulations and intermediate outcomes in COPD study (SPIROMICS). *Thorax* 2014;69:492–5.
26 Aaron CP, Hoffman EA, Kawut SM, *et al*. Ambient air pollution and pulmonary vascular volume on computed tomography: the MESA air pollution and lung cohort studies. *Eur Respir J* 2019;53:1802116.
27 Kaufman JD, Adar SD, Allen RW, *et al*. Prospective study of particulate air pollution exposures, subclinical Atherosclerosis, and clinical cardiovascular disease: the multi-ethnic study of Atherosclerosis and air pollution (MESA air). *Am J Epidemiol* 2012;176:825–37.
28 Regan EA, Hokanson JE, Murphy JR, *et al*. Genetic epidemiology of COPD (Copdgene) study design. *COPD* 2010;7:32–43.
29 Sieren JP, Newell JD, Barr RG, *et al*. SPIROMICS protocol for multicenter quantitative computed tomography to phenotype the lungs. *Am J Respir Crit Care Med* 2016;194:794–806.
30 Hoffman EA, Jiang R, Baumhauer H, *et al*. Reproducibility and validity of lung density measures from cardiac CT scans—the multi-ethnic study of Atherosclerosis (MESA) lung study. *Academic Radiology* 2009;16:689–99.
31 Gevenois PA, de Maertelaer V, De Vuyst P, *et al*. Comparison of computed density and macroscopic Morphometry in pulmonary emphysema. *Am J Respir Crit Care Med* 1995;152:653–7.
32 Hoffman EA, Ahmed FS, Baumhauer H, *et al*. Variation in the percent of emphysema-like lung in a healthy, nonsmoking multiethnic sample. The MESA lung study. *Ann Am Thorac Soc* 2014;11:898–907.
33 Gangeh MJ, Sorensen L, Shaker SB, *et al*. A Texton-based approach for the classification of lung parenchyma in CT images. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI); 2010:595–602
34 Gorelick L, Galun M, Sharon E, *et al*. Shape representation and classification using the Poisson equation. *IEEE Trans Pattern Anal Mach Intell* 2006;28:1991–2005.
35 Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci USA* 2008;105:1118–23.
36 Yang J, Angelini ED, Smith BM, *et al*. Explaining radiological emphysema subtypes with Unsupervised texture Prototypes: MESA COPD study. *Med Comput Vis Bayesian Graph Models Biomed Imaging (2016)* 2017;2017:69–80.
37 Häme Y, Angelini ED, Parikh ME, *et al*. Sparse sampling and unsupervised learning of lung texture patterns in pulmonary emphysema: MESA COPD study. *IEEE Int Symp Biomed Imaging* 2015:109–13.
38 Yang J, Angelini ED, Balte PP, *et al*. Unsupervised discovery of spatially-informed lung texture patterns for pulmonary emphysema: the MESA COPD study. *Med Image Comput Comput Assist Interv* 2017;10433:116–24.
39 Maaten LV, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9:2579–605.
40 Ganin Y, Lempitsky V. Unsupervised domain adaptation by Backpropagation. International Conference on Machine Learning; 2015:1180–9
41 Sack CS, Doney BC, Podolanczuk AJ, *et al*. Occupational exposures and subclinical interstitial lung disease. The MESA (multi-ethnic study of Atherosclerosis) air and lung studies. *Am J Respir Crit Care Med* 2017;196:1031–9.
42 Kim V, Davey A, Comellas AP, *et al*. Clinical and computed Tomographic predictors of chronic Bronchitis in COPD: a cross sectional analysis of the Copdgene study. *Respir Res* 2014;15:52.
43 Miller MR, Crapo R, Hankinson J, *et al*. General considerations for lung function testing. *Eur Respir J* 2005;26:153–61.
44 McDonough JE, Yuan R, Suzuki M, *et al*. Small-airway obstruction and emphysema in chronic obstructive pulmonary disease. *N Engl J Med* 2011;365:1567–75.
45 Galbán CJ, Han MK, Boes JL, *et al*. Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. *Nat Med* 2012;18:1711–5.
46 Smith BM, Kirby M, Hoffman EA, *et al*. Association of Dysanapsis with chronic obstructive pulmonary disease among older adults. *JAMA* 2020;323:2268–80.
47 Oelsner EC, Loehr LR, Henderson AG, *et al*. Classifying chronic lower respiratory disease events in epidemiologic cohort studies. *Ann Am Thorac Soc* 2016;13:1057–66.
48 The GTEx Consortium, Aguet F, Anand S, *et al*. The Gtex consortium Atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318–30.
49 Giambartolomei C, Vukcevic D, Schadt EE, *et al*. Bayesian test for Colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 2014;10:e1004383.
50 Wallace C. Eliciting Priors and relaxing the single causal variant assumption in Colocalisation analyses. *PLoS Genet* 2020;16:e1008720.
51 Herman AI, DeVito EE, Jensen KP, *et al*. Pharmacogenetics of nicotine addiction: role of dopamine. *Pharmacogenomics* 2014;15:221–34.
52 Matsuyama N, Shibata S, Matoba A, *et al*. The dopamine D_1 receptor is expressed and induces CREB Phosphorylation and MUC5AC expression in human airway epithelium. *Respir Res* 2018;19:53.
53 Kesimer M, Ford AA, Ceppe A, *et al*. Airway Mucin concentration as a marker of chronic Bronchitis. *N Engl J Med* 2017;377:911–22.
54 Kesimer M, Smith BM, Ceppe A, *et al*. Mucin concentrations and peripheral airways obstruction in COPD. *Am J Respir Crit Care Med* 2018.
55 Martinez FJ, Han MK, Allinson JP, *et al*. At the root: defining and halting progression of early chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2018;197:1540–51.
56 Burrows B, Kettel LJ, Niden AH, *et al*. Patterns of cardiovascular dysfunction in chronic obstructive pulmonary disease. *N Engl J Med* 1972;286:912–8.
57 Hueper K, Vogel-Claussen J, Parikh MA, *et al*. Pulmonary microvascular blood flow in mild chronic obstructive pulmonary disease and emphysema. The MESA COPD study. *Am J Respir Crit Care Med* 2015;192:570–80.
58 Thomashow MA, Shimbo D, Parikh MA, *et al*. Endothelial Microparticles in mild COPD and emphysema: the MESA COPD study. *Am J Respir Crit Care Med* 2013;188:60–8.
59 Wang M, Aaron CP, Madrigano J, *et al*. Association between long-term exposure to ambient air pollution and change in quantitatively assessed emphysema and lung function. *JAMA* 2019;322:546–56.
60 Auerbach O, Hammond EC, Garfinkel L, *et al*. Relationship of smoking and age to emphysema. whole-lung section study. *N Engl J Med* 1972;286:853–7.
61 Bickerman HA. Senile emphysema. *J Am Geriatr Soc* 1956;4:526–34.
62 Schiffers C, Lundblad LKA, Hristova M, *et al*. Downregulation of DUOX1 function contributes to aging-related impairment of innate airway injury responses and accelerated senile emphysema. *Am J Physiol Lung Cell Mol Physiol* 2021;321:L144–58.
63 Wicher SA, Roos BB, Teske JJ, *et al*. Aging increases Senescence, calcium signaling, and extracellular matrix deposition in human airway smooth muscle. *PLoS One* 2021;16:e0254710.
64 Cheng C-Y, Yamashiro K, Jia Chen L, *et al*. New Loci and coding variants confer risk for age-related macular degeneration in East Asians. *Nat Commun* 2015;6:6063.
65 Fishilevich S, Zimmerman S, Kohn A, *et al*. Genic insights from integrated human Proteomics in Genecards. *Database (Oxford)* 2016;2016:baw030.
66 Li Z-G, Scott MJ, Brzóska T, *et al*. Lung epithelial cell-derived IL-25 negatively regulates LPS-induced Exosome release from Macrophages. *Mil Med Res* 2018;5:24.
67 Srivastava A, Amreddy N, Razaq M, *et al*. Exosomes as Theranostics for lung cancer. *Adv Cancer Res* 2018;139:1–33.
68 Lin H, Jiang S. Combined pulmonary fibrosis and emphysema (CPFE): An entity different from emphysema or pulmonary fibrosis alone. *J Thorac Dis* 2015;7:767–79.
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 Ladizinski B, Sankey C. Vanishing lung syndrome. *N Engl J Med* 2014;370:e14.
71 Li F, Choi J, Zou C, *et al*. Latent traits of lung tissue patterns in former smokers derived by dual channel deep learning in computed tomography images. *Sci Rep* 2021;11.
72 Haghighi B, Choi S, Choi J, *et al*. Imaging-based clusters in former smokers of the COPD cohort associate with clinical characteristics: the subpopulations and intermediate outcome measures in COPD study (SPIROMICS). *Respir Res* 2019;20:153.
73 Binder P, Batmanghelich NK, Estépar RSJ, *et al*. Unsupervised discovery of emphysema subtypes in a large clinical cohort. In: *International Workshop on Machine Learning in Medical Imaging*. Springer, 2016: 180–7.