JCI insight

Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease

Sandeep Bodduluri,[1,2,3] Arie Nakhmani,[4] Joseph M. Reinhardt,[5] Carla G. Wilson,[6] Merry-Lynn McDonald,[2,3] Ramaraju Rudraraju,[7] Byron C. Jaeger,[8] Nirav R. Bhakta,[9] Peter J. Castaldi,[10] Frank C. Sciurba,[11] Chengcui Zhang,[12] Purushotham V. Bangalore,[12] and Surya P. Bhatt[1,2,3]

[1]UAB Lung Imaging Core, [2]UAB Lung Health Center, [3]Division of Pulmonary, Allergy and Critical Care Medicine, and [4]Department of Electrical and Computer Engineering, University of Alabama at Birmingham, Birmingham, Alabama, USA. [5]Department of Biomedical Engineering, University of Iowa, Iowa City, Iowa, USA. [6]Department of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado, USA. [7]Division of Cardiothoracic Surgery and [8]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, USA. [9]Division of Pulmonary, Critical Care, Allergy and Sleep Medicine, University California, San Francisco, San Francisco, California, USA. [10]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [11]Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [12]Department of Computer Science, University of Alabama at Birmingham, Birmingham, Alabama, USA.

BACKGROUND. Currently recommended traditional spirometry outputs do not reflect the relative contributions of emphysema and airway disease to airflow obstruction. We hypothesized that machine-learning algorithms can be trained on spirometry data to identify these structural phenotypes.

METHODS. Participants enrolled in a large multicenter study (COPDGene) were included. The data points from expiratory flow-volume curves were trained using a deep-learning model to predict structural phenotypes of chronic obstructive pulmonary disease (COPD) on CT, and results were compared with traditional spirometry metrics and an optimized random forest classifier. Area under the receiver operating characteristic curve (AUC) and weighted F-score were used to measure the discriminative accuracy of a fully convolutional neural network, random forest, and traditional spirometry metrics to phenotype CT as normal, emphysema-predominant (>5% emphysema), airway-predominant (Pi10 > median), and mixed phenotypes. Similar comparisons were made for the detection of functional small airway disease phenotype (>20% on parametric response mapping).

RESULTS. Among 8980 individuals, the neural network was more accurate in discriminating predominant emphysema/airway phenotypes (AUC 0.80, 95%CI 0.79–0.81) compared with traditional measures of spirometry, FEV_1/FVC (AUC 0.71, 95%CI 0.69–0.71), FEV_1% predicted (AUC 0.70, 95%CI 0.68–0.71), and random forest classifier (AUC 0.78, 95%CI 0.77–0.79). The neural network was also more accurate in discriminating predominant emphysema/small airway phenotypes (AUC 0.91, 95%CI 0.90–0.92) compared with FEV_1/FVC (AUC 0.80, 95%CI 0.78–0.82), FEV_1% predicted (AUC 0.83, 95%CI 0.80–0.84), and with comparable accuracy with random forest classifier (AUC 0.90, 95%CI 0.88–0.91).

CONCLUSIONS. Structural phenotypes of COPD can be identified from spirometry using deep-learning and machine-learning approaches, demonstrating their potential to identify individuals for targeted therapies.

TRIAL REGISTRATION. ClinicalTrials.gov NCT00608764.

FUNDING. This study was supported by NIH grants K23 HL133438 and R21EB027891 and an American Thoracic Foundation 2018 Unrestricted Research Grant. The COPDGene study is supported by NIH grants NHLBI U01 HL089897 and U01 HL089856. The COPDGene study

Conflict of interest: JMR reports holding stock in VIDA Diagnostics and personal fees from Boehringer Ingelheim. PJC reports personal fees and grant support from GlaxoSmithKline and personal fees from Novartis in the past 3 years. SPB has served on advisory boards for Sunovion and GlaxoSmithKline.

Submitted: August 16, 2019
Accepted: June 3, 2020
Published: June 18, 2020.

Reference information: *JCI Insight*. 2020;5(13):e132781. https://doi.org/10.1172/jci.insight.132781.

(NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprising AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Novartis, and Sunovion.

Introduction

Chronic obstructive pulmonary disease (COPD) is an inflammatory disease of the lungs that is associated with substantial respiratory morbidity and health care costs and is now the fourth leading cause of death in the United States (1). COPD is defined by persistent airflow obstruction on spirometry, the result of a combination of 2 distinct structural processes: emphysema characterized by alveolar destruction and poor elastic recoil of the lungs as well as airway disease characterized by airway narrowing and remodeling (2, 3). Although spirometric measures of airflow obstruction correlate strongly with CT measures of both emphysema and airway disease, spirometry does not discern the relative contributions of these structural disease processes to overall airflow obstruction. Furthermore, recent studies demonstrate that approximately half of current and former smokers, with no evidence of spirometric airflow obstruction according to traditional criteria, have evidence of emphysema and/or airway disease (4, 5). These findings suggest that the existing spirometry criteria for airflow obstruction are not sensitive to the contributory structural changes.

The inability to accurately and easily differentiate predominant emphysema from predominant airway disease hinders the development of targeted therapies (6). Furthermore, these structural changes have significant consequences beyond those due to lung function impairment. The degree of emphysema and airway wall thickening on CT are both independently associated with worse respiratory quality of life, dyspnea, and mortality (7–13). Despite these associations, CT is often not recommended for diagnosis in clinical practice due to concerns about high costs and risk of radiation. There are currently no low-cost, low-risk tools to phenotype the structural components of COPD, and even its diagnosis relies on demonstrating abnormalities in discrete components of spirometry, such as the forced expiratory volume in first second (FEV_1) and the ratio of FEV_1 to the forced vital capacity (FEV_1/FVC). Specific components of the spirometric flow-volume and volume-time curves have been analyzed to identify early and mild COPD but have not been successful in distinguishing emphysema-predominant disease from airway disease predominance (14–20).

We hypothesized that machine-learning approaches trained on all the data points contained in the expiratory flow-volume curve would accurately distinguish individuals with predominant emphysema from those with predominant airway disease. We used fully convolutional network (FCN) and random forest classifier to test our hypothesis.

Results

Participant characteristics

After exclusions (Figure 1), the expiratory flow-volume curves of 8980 participants were included in the analyses (Table 1). The mean (SD) age was 57.8 years (8.4 years); 4177 participants (39%) were female, 6085 participants (67.7%) were non-Hispanic white, and 2895 participants (32.3%) were African Americans. 4705 participants (52.9%) were active smokers at enrollment. The cohort included 3926 participants (44.1%) without airflow obstruction (GOLD 0) and 3901 participants (43.4%) with airflow obstruction, including 724 (8.1%), 1696 (19.1%), 984 (11.1%), and 497 (5.6%) with GOLD stages 1–4, respectively. 1066 participants (11.9%) had preserved ratio impaired spirometry (PRISm).

Structurally normal CT scans were seen in 3085 participants (34.3%). Airway-predominant disease, defined by Pi10, was seen in 3207 participants (35.7%), emphysema-predominant disease was seen in 1390 participants (15.4%), and a mixed phenotype was seen in 1298 participants (14.4%) (Table 1). Airway-predominant disease, defined by functional small airway disease, was seen in 826 participants (10.25%), emphysema-predominant disease was seen in 200 participants (2.54%), and a mixed phenotype was seen in 1636 participants (20.79%).

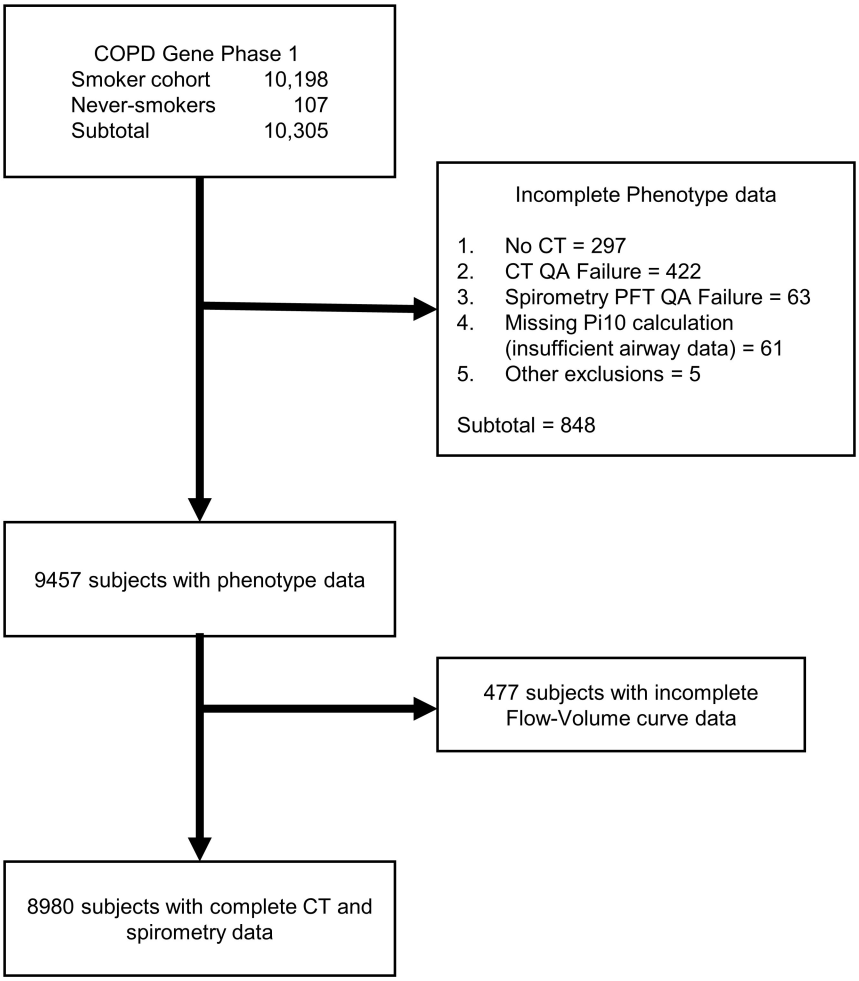
COPD Gene Phase 1
Smoker cohort 10,198
Never-smokers 107
Subtotal 10,305
Incomplete Phenotype data
1. No CT = 297
2. CT QA Failure = 422
3. Spirometry PFT QA Failure = 63
4. Missing Pi10 calculation (insufficient airway data) = 61
5. Other exclusions = 5
Subtotal = 848
9457 subjects with phenotype data
477 subjects with incomplete Flow-Volume curve data
8980 subjects with complete CT and spirometry data

Figure 1. CONSORT diagram.

Classification results for phenotyping emphysema/airway disease

Training. Results presented here are aggregated means from 10 replications of Monte-Carlo cross-validation in the training data set. The average out-of-the-bag error for the random forest model was 0.45 (95%CI 0.44 to 0.46), whereas the average validation loss for the neural network was 1.01 (95%CI 1.00 to 1.02). The parameters and weights of the model with minimum validation loss (neural network) and minimum out-of-the-bag error (random forest) among the 10 training/validation splits were used to evaluate the model performance on the held-out test set.

Held-out test data set. In the test data (20% of cohort), (1796 participants), 617 participants (34.3%) were normal, 641 participants (35.6%) had predominant airway disease, 278 participants (15.4%) had emphysema predominant disease, and 260 participants (14.4%) participants had a mixed phenotype. For the prediction of structural phenotypes, the AUCs for FEV_1% predicted and FEV_1/FVC were 0.70 (95%CI 0.68 to 0.71) and 0.71 (95%CI 0.69 to 0.71), respectively. The area under the receiver operating characteristic curve (AUC) for the random forest classification was 0.78 (95%CI 0.77 to 0.79). The neural network outperformed traditional measures of spirometry and also the optimized random forest classifier with AUC of 0.80 (95%CI 0.79 to 0.81) (Table 2). The F1 score for the neural network was 0.56 compared with 0.45, 0.43, and 0.54 for FEV_1% predicted, FEV_1/FVC, and random forest classifier, respectively (Figure 2 and Figure 3). Within each structural disease class, neural network again outperformed the traditional spirometry measures for classification (Table 2 and Figures 2 and 3). Results for feature importance using SHapley Additive exPlanation (SHAP) values are shown in Supplemental Figures 1 and 2 (supplemental material available online with this article; https://doi.org/10.1172/jci.insight.132781DS1).

Table 1. Baseline characteristics of participants in each structural phenotype

	Normal	Emphysema predominant	Airway disease predominant	Mixed phenotype
No. of subjects	3085	1390	3207	1298
Age, years	57.8 (8.4)	63.1 (8.2)	57.3 (8.5)	65.5 (8.1)
Sex, *n* (%)				
Female	1210 (39.2%)	438 (31.5%)	1938 (60.4%)	591 (45.5%)
Male	1875 (60.8%)	952 (68.5%)	1269 (29.6%)	707 (54.5%)
Race/ethnicity, *n* (%)				
Non-Hispanic White	2114 (68.6%)	1160 (83.5%)	1766 (55.1%)	1045 (80.6%)
Non-Hispanic Black	971 (31.4%)	230 (16.5%)	1441 (44.9%)	253 (19.4%)
BMI (kg/m^2)	28.5 (5.3)	27.2 (5.6)	30.2 (6.8)	26.6 (5.6)
Smoking pack-years	39.8 (21.2)	49.3 (26.3)	41.4 (23.4)	55.3 (29.9)
Current smokers, n (%)	1757 (56.9%)	413 (29.7%)	2155 (67.1%)	380 (29.2%)
FEV$_1$, L	2.8 (0.7)	2.1 (0.9)	2.2 (0.7)	1.2 (0.7)
FEV$_1$% predicted	90.9 (16.3)	68.9 (27.1)	80.2 (20.0)	47.0 (23.7)
FEV$_1$/FVC	0.74 (0.0)	0.56 (0.1)	0.73 (0.1)	0.45 (0.1)
GOLD stage, *n* (%)				
Nonsmokers	58 (66.6%)	5 (0.05%)	21 (24.1%)	3 (0.03%)
PRISm	313 (29.3%)	39 (0.03%)	689 (64.6%)	25 (0.02%)
GOLD 0	2018 (51.4%)	328 (8.3%)	1499 (38.1%)	81 (2.0%)
GOLD 1	295 (40.7%)	203 (28.0%)	173 (23.8%)	53 (7.3%)
GOLD 2	368 (21.6%)	417 (24.5%)	585 (34.4%)	326 (19.2%)
GOLD 3	29 (2.9%)	281 (28.5%)	210 (21.3%)	464 (47.1%)
GOLD 4	4 (0.08%)	117 (23.5%)	30 (6.0%)	346 (69.6%)
Percentage emphysema on CT	1.6 (1.3)	15.3 (10.9)	1.1 (1.2)	18.7 (11.7)
Pi10	3.5 (0.0)	3.5 (0.0)	3.7 (0.1)	3.7 (0.1)

All values are expressed as mean (SD) unless specified otherwise. FEV$_1$, forced expiratory volume in the first second; FVC, forced vital capacity; GOLD, Global Initiative for Chronic Obstructive Lung Disease; PRISm, preserved ratio impaired spirometry; Pi10, square root of wall area of a theoretical airway with 10-mm luminal perimeter.

Classification results for phenotyping emphysema/functional small airway disease

Training. The average out-of-the-bag error for the random forest model was 0.19 (95%CI 0.20 to 0.18), whereas the average validation loss for the neural network was 0.57 (95%CI 0.55 to 0.59).

Held out test data set. In the test data (20% of cohort), (1574 participants), 1041 (66.1%) were normal, 165 (10.4%) had predominant small airway disease, 40 (2.5%) had emphysema predominant disease, and 328 (20.8%) participants had a mixed phenotype. For the prediction of structural phenotypes, the AUCs for FEV$_1$% predicted and FEV$_1$/FVC were 0.83 (95%CI 0.80 to 0.84) and 0.80 (95%CI0.78 to 0.82), respectively. The AUC for the random forest classification was 0.90 (95%CI 0.88 to 0.91). The neural network outperformed traditional measures of spirometry and had similar discrimination compared with the random forest classifier with AUC of 0.91 (95%CI 0.90 to 0.92) (Table 3). The F1 score for the neural network was 0.79 compared with 0.73, 0.71, and 0.76 for FEV$_1$% predicted, FEV$_1$/FVC, and random forest classifier, respectively (Figure 4 and Figure 5). Within each structural disease class, FCN again outperformed the traditional spirometry measures for classification (Table 3 and Figures 4 and 5).

Discussion

In a large multicenter cohort of current and former smokers, machine-learning approaches, including deep-learning methods, trained on spirometry data outperformed traditional spirometry measures for the phenotyping of COPD into its structural components, including predominant small airway disease, and provided flow-volume curve signatures for predominant structural disease categories. These results will enhance patient identification for phenotypic characterization and targeting therapies.

Spirometric impairment is a summary metric, and the development of targeted therapies is hindered by the inability to identify predominant COPD phenotypes. Existing threshold-based spirometry criteria are also insensitive to early and mild damage in the lungs. As much as 20%–25% of the lung may be affected by emphysema before these changes manifest on spirometry (21). Substantial airway remodeling and loss also

Table 2. Discriminative accuracy of traditional spirometry metrics, random forest classifier, and deep-learning model for emphysema/medium size airway disease

		Normal	Airway disease	Emphysema	Mixed
FEV$_1$/FVC	AUC (95%CI)	0.70 (0.67, 0.72)	0.63 (0.61, 0.66)	0.73 (0.71, 0.75)	0.89 (0.88, 0.90)
	Δ AUC[A](95%CI)	−0.10 (−0.12, −0.07)	−0.15 (−0.17, −0.12)	−0.05 (−0.06, −0.01)	−0.01 (−0.02, −0.01)
	Sensitivity (95%CI)	0.90 (0.88, 0.93)	0.92 (0.90, 0.95)	0.69 (0.63, 0.74)	0.85 (0.80, 0.89)
	Specificity (95%CI)	0.47 (0.44, 0.50)	0.30 (0.28, 0.33)	0.68 (0.66, 0.70)	0.84 (0.82, 0.86)
	Youden index (95%CI)	0.37 (0.33, 0.40)	0.23 (0.19, 0.26)	0.37 (0.30, 0.42)	0.68 (0.63, 0.72)
FEV$_1$% predicted	AUC (95%CI)	0.74 (0.72, 0.76)	0.63 (0.58, 0.62)	0.60 (0.58, 0.62)	0.87 (0.84, 0.89)
	Δ AUC[A] (95%CI)	−0.06 (−0.07, −0.03)	−0.15 (−0.12, −0.18)	−0.18 (−0.20, −0.13)	−0.03 (−0.05, −0.02)
	Sensitivity (95%CI)	0.82 (0.80, 0.86)	0.66 (0.63, 0.70)	0.46 (0.40, 0.52)	0.80 (0.74, 0.84)
	Specificity (95%CI)	0.56 (0.54, 0.60)	0.53 (0.50, 0.56)	0.72 (0.70, 0.74)	0.82 (0.80, 0.84)
	Youden index (95%CI)	0.39 (0.35, 0.42)	0.19 (0.14, 0.22)	0.18 (0.11, 0.23)	0.61 (0.55, 0.66)
Random forest	AUC (95%CI)	0.78 (0.76, 0.80)	0.75 (0.75, 0.79)	0.75 (0.73, 0.77)	0.90 (0.88, 0.92)
	Δ AUC[A] (95%CI)	−0.01 (−0.02, −0.00)	−0.02 (−0.03, −0.01)	−0.02 (−0.03, −0.01)	−0.002 (−0.009, −0.004)
	Sensitivity (95%CI)	0.83 (0.80, 0.86)	0.74 (0.70, 0.77)	0.76 (0.70, 0.81)	0.87 (0.82, 0.91)
	Specificity (95%CI)	0.60 (0.58, 0.63)	0.64 (0.61, 0.67)	0.63 (0.61, 0.66)	0.81 (0.79, 0.83)
	Youden index (95%CI)	0.43 (0.40, 0.46)	0.38 (0.33, 0.41)	0.38 (0.32, 0.42)	0.68 (0.62, 0.72)
Neural network	AUC (95%CI)	0.80 (0.78, 0.81)	0.78 (0.75, 0.79)	0.78 (0.76, 0.79)	0.91 (0.89, 0.92)
	Δ AUC	Reference	Reference	Reference	Reference
	Sensitivity (95%CI)	0.72 (0.68, 0.76)	0.76 (0.73, 0.80)	0.62 (0.56, 0.68)	0.89 (0.84, 0.92)
	Specificity (95%CI)	0.74 (0.72, 0.77)	0.67 (0.64, 0.70)	0.79 (0.77, 0.81)	0.81 (0.79, 0.83)
	Youden index (95%CI)	0.46 (0.41, 0.49)	0.43 (0.38, 0.46)	0.41 (0.31, 0.45)	0.70 (0.65, 0.73)

FEV$_1$, forced expiratory volume in the first second; FVC, forced vital capacity; AUC, area under the curve. [A]Comparison of AUC using DeLong method with FCN as reference model.

occur before the development of significant spirometric impairment (22, 23). Because airflow in the middle part of the flow-volume curve disproportionately results from flow in the small airways, multiple studies have evaluated various methods of analyzing the mid and distal parts of the curve. These include the forced expiratory flow in the 25th to the 75th percentile, forced expiratory flow in the first 3 seconds, the shape of the maximum expiratory curve, and change in angle of flow during forced exhalation (14–20, 24–29). None of these studies, however, validated their measures against structural lung disease and, hence, were unable to separate emphysema from airway predominant disease.

Distinguishing predominant emphysema and airway disease is relevant for optimizing and advancing clinical care. Current therapy for COPD includes bronchodilators and inhaled corticosteroids, and only half of patients treated with these medications have a clinically meaningful improvement in their respiratory quality of life (30). These therapies target airway tone and inflammation and do not target emphysema. Although no specific pharmacologic therapies are currently approved that separately target emphysema and airway disease, interventional and pharmacologic therapies are being developed that are likely to benefit carefully phenotyped and selected individuals. Surgical and bronchoscopic lung volume reduction procedures are approved for severe emphysema. New interventions that target chronic bronchitis and airway remodeling are being developed, and there are ongoing trials that specifically target emphysema (NCT02696564). The results of this study can help identify these patients for clinical trials and eventually therapy.

Several aspects related to deep learning that are pertinent to this study should be considered. The prediction of structural lung disease from a sequence of flow values derived from a forced expiratory effort is effectively a sequence classification task. Capitalizing on recent advances in deep learning, several studies implemented convolutional neural networks and long short-term memory models to classify sequential data from natural language processing and speech recognition tasks. In the current study, we applied a fully convolutional network (FCN) as well as a random forest classifier on flow sequence generated from the expiratory flow-volume curve to phenotype structural lung disease as the outcome. Wang et al. proposed the use of FCNs to analyze sequential data and achieved robust results on several data sets from the University of California, Riverside, time series classification archive (31, 32). The FCN architecture was initially proposed for semantic image segmentation tasks, where the architecture is composed of 3 computation

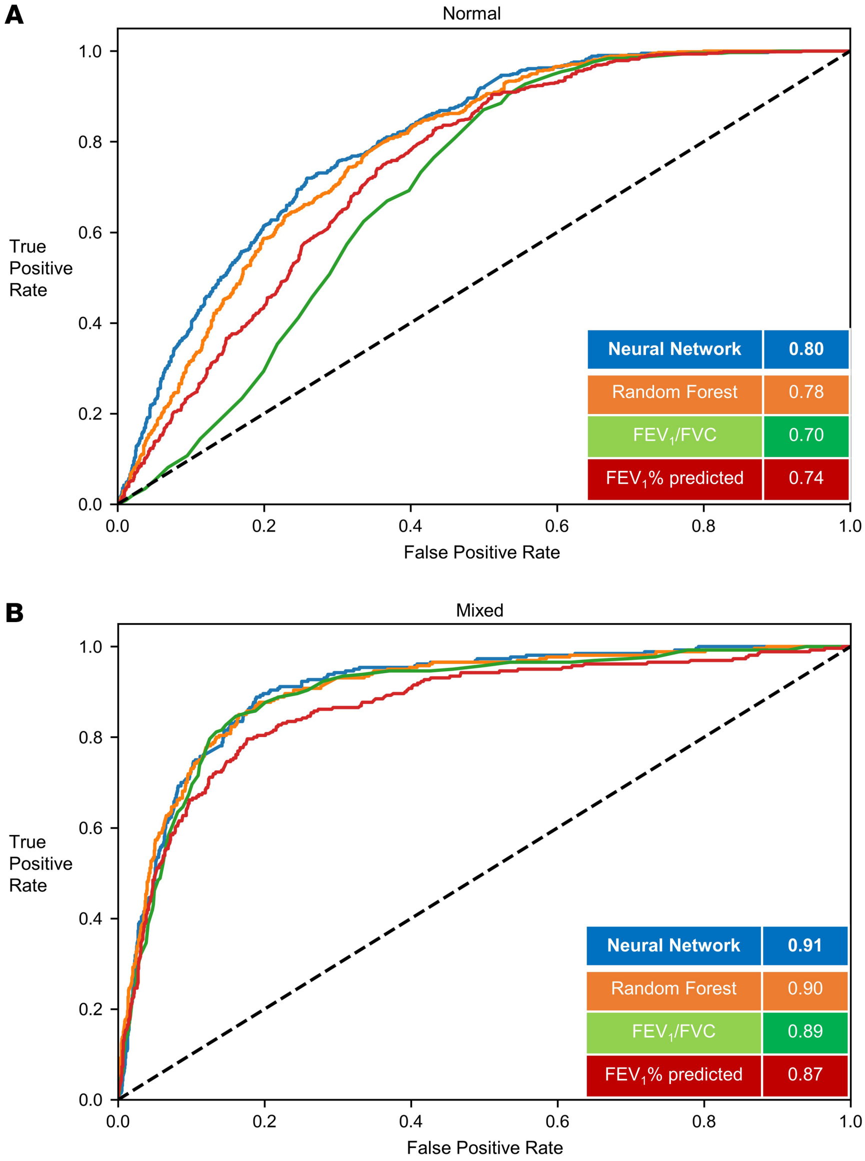

Figure 2. Classification of structural phenotypes — normal and mixed emphysema/airway disease. Classification performance of normal (**A**) (emphysema <5% and medium size airway disease < median Pi10) and mixed (**B**) (emphysema >5% and medium size airway disease > median Pi10) groups. The results show per-class area under the curve (AUC) of the FCN model versus random forest classifier and logistic regression models with FEV$_1$/FVC and FEV$_1$% predicted measurements. Results shown for the hold-out test data set. FEV$_1$, forced expiratory volume in the first second; FVC, forced vital capacity; FCN, fully convolutional network; Pi10, airway wall area measurement.

blocks and each block performs convolution operations followed by batch normalization and ReLU activation layers. The resulting output from the 3 convolution operations is fed into a global average pooling layer, which drastically reduces the number of training parameters and further enables the visualization of class activations specific to each class. The minimal requirement for preprocessing and feature crafting before classification, and the visualization of feature activations specific to each class through the pooling layer, make FCN an effective choice for classification of sequential or time series data in the medical domain. Nonetheless, a random forest classifier with optimized parameters performed almost equally well as the neural network, suggesting that a number of other machine-learning and deep-learning algorithms may be applied to the sequence of raw data points that constitute the spirometry curves. Random forest relies on decision points and avoids correlated points, whereas FCN uses near neighbors that may be correlated.

Although further improvement in accuracy may be possible with other algorithms, the overall results reflect the frequently observed overlapping and interrelated structural changes in both airways and the

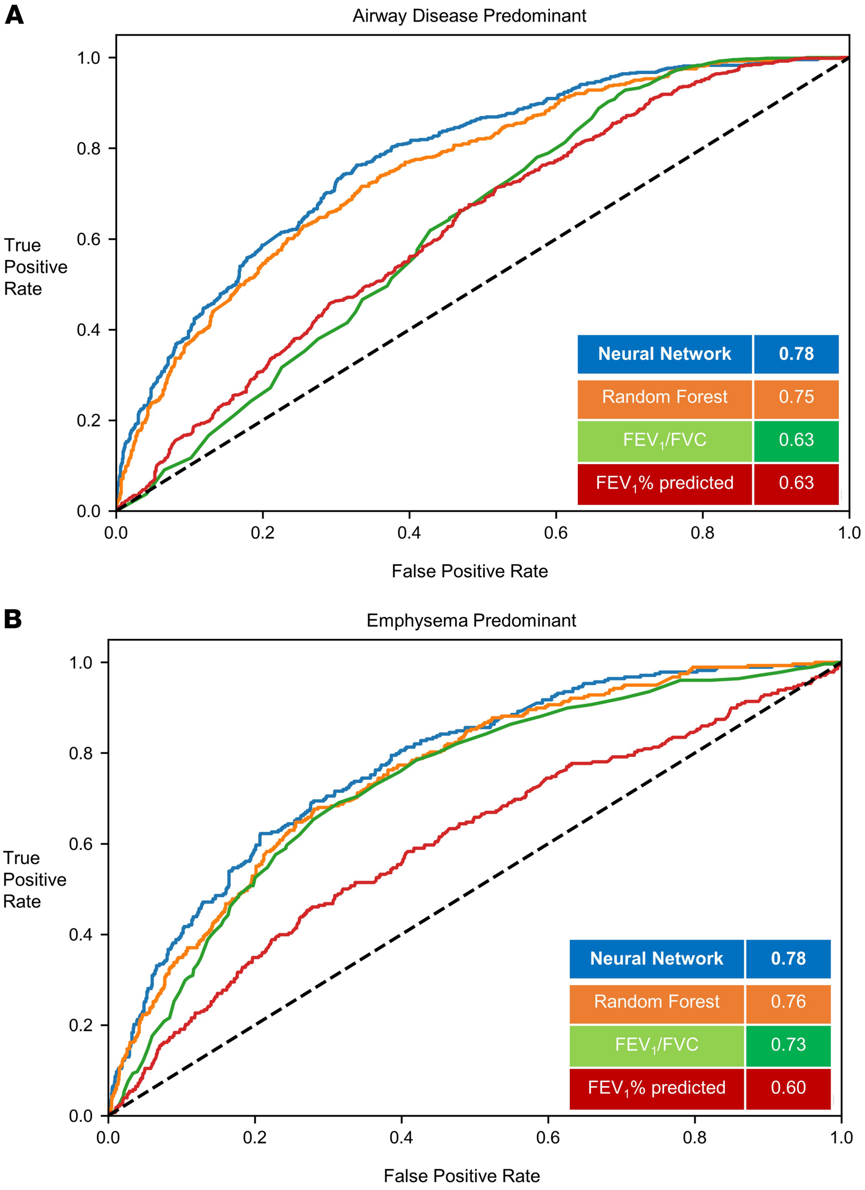

Figure 3. Classification of structural phenotypes — emphysema and airway disease. Classification performance of airway disease predominant (**A**) (emphysema <5% and medium size airway disease > median Pi10) and emphysema predominant (**B**) (emphysema >5% and medium size airway disease < median Pi10) groups. The results show per-class area under the curve (AUC) of the FCN model versus random forest classifier and logistic regression models with FEV_1/FVC and FEV_1% predicted measurements. Results shown for the hold-out test data set. FEV_1, forced expiratory volume in the first second; FVC, forced vital capacity; FCN, fully convolutional network; Pi10, airway wall area measurement.

parenchyma in varying proportions that occurs in the majority of smokers. There is considerable airway-parenchymal interdependence; the presence of emphysema can untether airways and result in a predisposition to airway collapse, and peribronchial fibrosis and airway loss can result in distal emphysema. Although this inherent biological complexity limits the information that can be ascertained from spirometry alone, the probability scores that result from the FCN model for each individual raise the likelihood of accurately identifying the predominant structural category and represent a substantial advance in the identification of structural phenotypes in COPD.

The study has several strengths. Data from a large multicenter cohort of participants whose disease spanned the range of severity were included. Extensive CT phenotyping was performed with stringent quality control of both CT and spirometry. The structural phenotypes were classified using quantitative CT data

Table 3. Discriminative accuracy of traditional spirometry metrics, random forest classifier, and deep-learning model for emphysema/functional small airway disease

		Normal	Airway disease	Emphysema	Mixed
FEV_1/FVC	AUC (95%CI)	0.82 (0.79, 0.84)	0.58 (0.53, 0.61)	0.59 (0.51, 0.66)	0.88 (0.85, 0.89)
	Δ AUC[A] (95%CI)	−0.10 (−0.12, −0.08)	−0.19 (−0.23, −0.15)	−0.21 (−0.30, −0.12)	−0.07 (−0.09, −0.06)
	Sensitivity (95%CI)	0.84 (0.81, 0.86)	0.82 (0.75, 0.87)	0.75 (0.59, 0.87)	0.74 (0.69, 0.78)
	Specificity (95%CI)	0.69 (0.65, 0.73)	0.32 (0.30, 0.35)	0.51 (0.48, 0.53)	0.87 (0.85, 0.89)
	Youden index (95%CI)	0.52 (0.47, 0.56)	0.14 (0.06, 0.18)	0.25 (0.11, 0.36)	0.61 (0.56, 0.65)
FEV_1% predicted	AUC (95%CI)	0.85 (0.82, 0.86)	0.61 (0.57, 0.64)	0.67 (0.58, 0.74)	0.90 (0.87, 0.91)
	Δ AUC[A] (95%CI)	−0.08 (−0.09, −0.06)	−0.16 (−0.19, −0.12)	−0.14 (−0.23, −0.04)	−0.06 (−0.07, −0.04)
	Sensitivity (95%CI)	0.90 (0.88, 0.91)	0.59 (0.52, 0.67)	0.70 (0.54, 0.83)	0.82 (0.78, 0.86)
	Specificity (95%CI)	0.67 (0.63, 0.71)	0.63 (0.60, 0.66)	0.66 (0.63, 0.68)	0.85 (0.83, 0.87)
	Youden index (95%CI)	0.56 (0.51, 0.60)	0.22 (0.13, 0.28)	0.35 (0.17, 0.47)	0.67 (0.62, 0.71)
Random forest	AUC (95%CI)	0.92 (0.90, 0.92)	0.73 (0.70, 0.76)	0.72 (0.65, 0.77)	0.95 (0.93, 0.95)
	Δ AUC[A] (95%CI)	−0.01 (−0.01, −0.00)	−0.04 (−0.06, −0.01)	−0.08 (−0.14, −0.02)	−0.009 (−0.01, −0.004)
	Sensitivity (95%CI)	0.90 (0.88, 0.92)	0.75 (0.67, 0.81)	0.75 (0.59, 0.87)	0.89 (0.85, 0.92)
	Specificity (95%CI)	0.80 (0.76, 0.83)	0.64 (0.61, 0.67)	0.63 (0.60, 0.66)	0.88 (0.86, 0.90)
	Youden index (95%CI)	0.69 (0.65, 0.72)	0.38 (0.30, 0.43)	0.37 (0.21, 0.46)	0.76 (0.72, 0.79)
Neural network	AUC (95%CI)	0.93 (0.91, 0.93)	0.77 (0.73, 0.80)	0.81 (0.75, 0.85)	0.96 (0.94, 0.96)
	Δ AUC	Reference	Reference	Reference	Reference
	Sensitivity (95%CI)	0.87 (0.85, 0.90)	0.72 (0.64, 0.78)	0.75 (0.58, 0.87)	0.94 (0.90, 0.96)
	Specificity (95%CI)	0.85 (0.82, 0.88)	0.70 (0.68, 0.73)	0.72 (0.70, 0.74)	0.84 (0.82, 0.86)
	Youden index (95%CI)	0.72 (0.68, 0.75)	0.41 (0.33, 0.46)	0.47 (0.32, 0.55)	0.77 (0.73, 0.80)

FEV_1, forced expiratory volume in the first second; FVC, forced vital capacity; AUC, area under the curve. [A]Comparison of AUC using DeLong method with FCN as reference model.

that are more objective than labels applied by experts and are less subject to variability. COPDGene (Genetic Epidemiology of COPD) included a substantial number of African Americans and women. The training of the neural network and subsequent hyperparameter optimization was performed over 10 replications of Monte-Carlo cross-validation to ensure robustness of the model. The final evaluation of the classifier was performed on a hold-out test data set, which was not seen by the model previously.

Limitations. The study also has several limitations. First, COPDGene included current and former smokers, and hence, these results should be validated in cohorts that include nonsmokers with and at risk for COPD. Second, CT scans were not spirometry gated. Participants were, however, coached to reproducibly achieve maximal inhalation. Third, the outcome variables were numeric values for CT parameters, and it is not known how factors that cause variability in CT assessments, such as scanner type and field of view, effect the performance of the machine-learning models. These aspects need further analysis. Fourth, although the top 5 flows at a given volume that are associated with each of the phenotypic classes were identified, we are unable to ascribe a physiologic explanation to these findings. The flow values in combination appear to reflect structural processes that are not detected when discrete single values are used. Although we used SHAP to identify the top features, interpretation is limited, as these features may slightly differ in a different data set, and hence, the inherent black box nature of FCN remains (33). Fifth, machine-learning algorithms can be affected by underfitting and overfitting biases (34), but we obtained similar results in a hold-out test data set.

Conclusions. Structural phenotypes of COPD can be identified from spirometry using a deep neural network and machine-learning approaches, demonstrating their potential to identify individuals for targeted therapies. Further research is necessary to evaluate the applicability of the deep-learning model to improve COPD outcomes.

Methods

Study population and physiologic assessments. Spirometry data from participants enrolled in the COPDGene study were included (35). COPDGene is a large multicenter cohort study of current and former smokers aged between 45 and 80 years, with a smoking history of at least 10-pack years; the details of this study have been previously published.

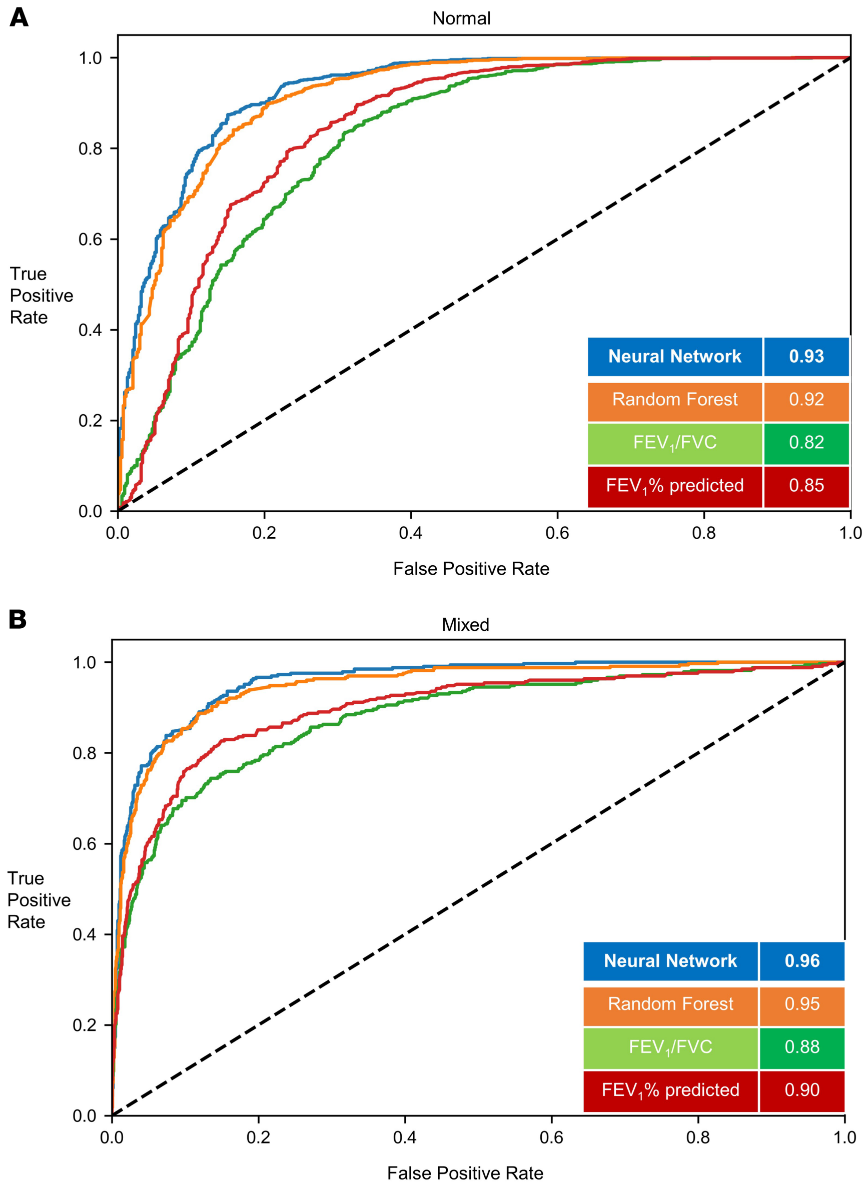

Figure 4. Classification of structural phenotypes — normal and mixed emphysema/small airway disease. Classification performance of normal (**A**) (emphysema <5% and small airway disease <20%) and mixed (**B**) (emphysema >5% and small airway disease >20%) groups. The results show per-class area under the curve (AUC) of the FCN model versus random forest classifier and logistic regression models with FEV_1/FVC and FEV_1% predicted measurements. Results shown for the hold-out test data set. FEV_1, forced expiratory volume in the first second; FVC, forced vital capacity; FCN, fully convolutional network. Small airway disease was defined by parametric response mapping.

All participants underwent a standard protocol, which included prebronchodilator and postbronchodilator spirometry using the New Diagnostic Design Easy-One spirometer per the American Thoracic Society criteria. Postbronchodilator spirometry was performed 20 minutes after administration of 180 μg albuterol HFA with a spacer (Aerochamber, Monaghan Medical Corporation). Quality control was performed by including only those spirometry efforts that met at least grade 2 ATS standards (repeatable between 100 and 150 ml). The postbronchodilator ratio of FEV_1/FVC < 0.70 was used to confirm the presence of airflow obstruction (36), and FEV_1% predicted was used to estimate the severity of airflow obstruction per Global initiative for Obstructive Lung Disease (GOLD) recommendations (37). Participants with FEV_1/FVC >0.70 but with FEV_1% predicted <80% were categorized as having PRISm (38). We selected the postbronchodilator effort with the highest sum of FEV_1 and FVC for the analysis as per ATS criteria. The raw data points that constitute the expiratory flow-volume curve were decomposed incrementally as flow data at every 30 mL volume exhaled (39, 40).

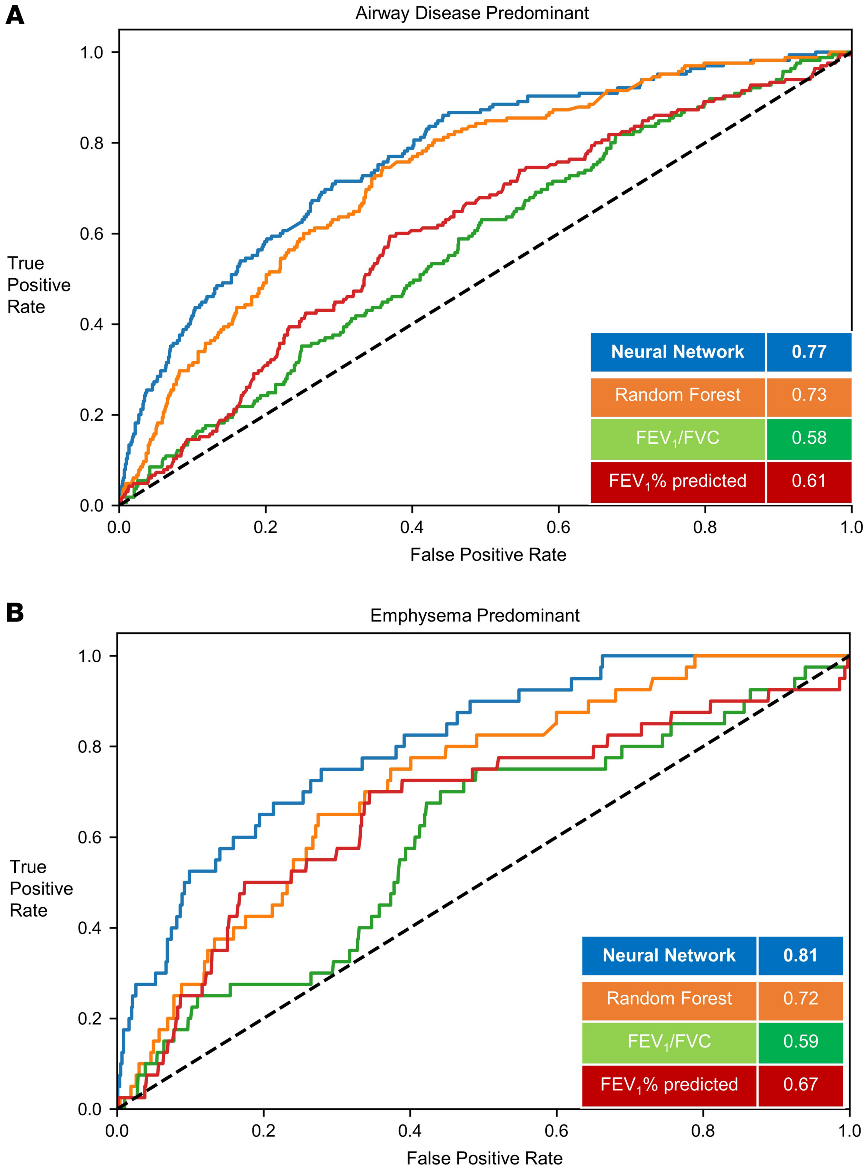

Figure 5. Classification of structural phenotypes — emphysema and small airway disease. Classification performance of airway disease predominant (**A**) (emphysema <5% and small airway disease >20%) and emphysema predominant (**B**) (emphysema >5% and small airway disease <20%) groups. The results show per-class area under the curve (AUC) of the FCN model versus random forest classifier and logistic regression models with FEV_1/FVC and FEV_1% predicted measurements. Results shown for the hold-out test data set. FEV_1, forced expiratory volume in the first second; FVC, forced vital capacity; FCN, fully convolutional network. Small airway disease was defined by parametric response mapping.

CT-based phenotyping. Quantitative CT scans were acquired at maximal inspiration (total lung capacity). Emphysema was quantified on inspiratory CT as the percentage of low attenuation areas <–950 Hounsfield units using Slicer 3D software (11). Clinically significant emphysema was defined as ≥5% low attenuation areas. This threshold was selected as there appears to be an inflection point at 5%, above which the frequency of exacerbations and mortality increases considerably (12). Large and medium size airway disease was quantified by the Pi10, the square root of the wall area of a hypothetical airway with internal perimeter of 10 mm, using Apollo Software (VIDA Diagnostics) (11). Because there is no established threshold for clinically significant airway wall disease, Pi10 > median in the COPDGene cohort was used for categorization as significant airway disease. Functional small airway disease (fSAD) phenotype was quantified by >20% lung affected by small airway disease measured on parametric response mapping, where fSAD is nonemphysematous air trapping and, hence, an indirect measure of small airway disease (41, 42).

Using these emphysema and Pi10 thresholds, we classified participants into 1 of 4 CT categories: normal, <5% emphysema and < median Pi10; airway predominant, <5% emphysema but with Pi10 ≥ median; emphysema predominant, ≥5% emphysema and Pi10 < median; and mixed emphysema/airway, ≥5% emphysema and ≥ median Pi10. We also classified participants into 4 groups based on fSAD: normal, <5% emphysema and <20% fSAD; airway predominant, <5% emphysema but with fSAD ≥20%; emphysema predominant, ≥5% emphysema and fSAD <20%; and mixed emphysema/airway, ≥5% emphysema and fSAD ≥20%.

Deep neural network. FCN was developed for image segmentation tasks and has shown significant improvements in efficiency and overall performance as compared with traditional deep convolutional networks. In this study, we used FCN as a feature extractor of a time series (or sequential) data, where these features were further fed into a global average pooling layer and a soft-max layer to classify the sequences into different labels. The basic architecture of FCN includes 3 stacked computation blocks, where each block consists of 1D convolutional layer followed by a batch normalization layer and a rectified linear unit activation layer (Supplemental Figure 3). Convolution on the 1D input sequence was performed by the convolutional layers followed by the batch normalization layer to improve generalizability and faster convergence. The penultimate global average pooling layer reduces the number of weights and prevents overfitting. This FCN architecture has been previously shown to achieve superior performance in several 1D sequence classification tasks.

Model training and evaluation. The flow data points in each expiratory flow-volume curve were used as a 1D input sequence, and each sequence was standardized to have a length of 200 points using data padding with zeros at the end of the sequence. The expiratory flow data was divided into input (80%) and hold-out test (20%) data sets. All possible combinations of number of filters (32, 64, 128, 256 filters) in the convolutional layers, learning rate in the range of 0.00001–0.1, and batch sizes of 64, 128, and 256 were evaluated on the training set to select the best hyperparameters. The hyperparameter tuning was performed using TALOS library in Python. The model with the best hyperparameters, where the 3 convolutional layers with filter sizes of 128, 256, and 128, corresponding kernel sizes of 9, 5, and 3, at a learning rate of 0.0001, with batch size of 64 over 100 epochs, was selected for further evaluation. The input data set was further divided into 10 random splits of training (80%) and validation (20%) to train the FCN model. The weights of the neural network with minimum loss on the validation set were used for subsequent evaluation on the hold-out test data set. Early stopping of the training was implemented when there was no decline in the validation loss for at least 25 epochs. The learning rate was reduced by a factor of 0.01 after 15 epochs of no decline in the validation loss. The primary outcome was classification of each participant into 1 of the 4 structural disease categories on quantitative CT. Supplemental Figure 4 shows the visualization of the FCN training process with the chosen hyperparameters to classify spirometry data into the 4 different structural COPD phenotypes. The performance of the FCN was compared by implementing optimized random forest model (parameters were chosen by 5-fold cross validation and selected the model with minimum out-of-the-bag error) on the same input sequences and also with the performance of the traditional spirometry variables (FEV_1/FVC and FEV_1% predicted). Computation of feature importance using SHAP values is described in the Supplemental Methods.

Statistics. AUC analyses were computed to evaluate the accuracy of the FCN and the random forest classifier. Their discriminative accuracies were compared with 2 traditional spirometry measurements (FEV_1/FVC and FEV_1% predicted) based on logistic regression. Sensitivity, specificity, Youden index (sensitivity + specificity −1), and F1 score for structural disease classification were tested for each model (43). The nonparametric DeLong test was used to compare AUCs between the models (44). A 2-tailed *P* value of < 0.05 was considered significant for all analyses. Analyses were performed using Python ≥ 3.0, R version ≥ 3.6.0 (R Project for Statistical Computing), and MedCalc Statistical Software.

Study approval. All participants provided written informed consent before enrollment, and the COPDGene study protocol was approved by the University of Alabama at Birmingham Institutional Review Board (IRB) for human use (F070712014). The COPDGene study was approved by the IRBs of all 21 participating clinical centers: Ann Arbor VA Medical Center IRB (no. 2014-060462), Ann Arbor, Michigan, USA; Baylor College of Medicine IRB (H-22209); Brigham and Women's Hospital Partners Human Research Committee (no. 2007P000554); Columbia University IRB (AAAC9324), New York, New York, USA; Duke University Health System IRB (no. Pro00004464), Durham, North Carolina, USA; Johns Hopkins Medicine IRB (NA_00011524), Baltimore, Maryland, USA; Los Angeles Biomedical Research Institute Human Subjects Committee (no. 12756-03), Torrance, California, USA; Michael E. DeBakey VA Medical

Center IRB (no. H-22202), Houston, Texas, USA; Minneapolis VA Health Care System Minnesota (no. 4128-A), Minneapolis, Minnesota, USA; Health Partners Twin Cities IRB, (no. 07-127) Minneapolis–Saint Paul, Minnesota, USA; Morehouse School of Medicine IRB (no. 97826), Atlanta, Georgia, USA; National Jewish Health IRB (no. 1883a); Reliant Medical Group IRB (Fallon) (no. 1441), Worcester, Massachusetts, USA; Temple University IRB (no. 21659), Philadelphia, Pennsylvania, USA; University California, San Diego, Human Research Protections Program (no. 140070), San Diego, California, USA; University of Iowa IRB (no. 200710717); University of Michigan Medical School IRB (HUM00014973), Ann Arbor, Michigan, USA; University of Minnesota IRB Human Subjects Committee (no. 0801M24949), Minneapolis, Minnesota, USA; University of Pittsburgh IRB (no. 07120059); and University of Texas Health Science Center at San Antonio IRB (HSC20070644H), San Antonio, Texas, USA.

Author contributions

SB and SPB had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. SB and SPB conceived and designed the study. SB, AN, JMR, CGW, MLM, RR, BCJ, NRB, PJC, FCS, CZ, PVB, SPB acquired, analyzed, or interpreted data. SB and SPB drafted the manuscript. SB, AN, JMR, CGW, MLM, RR, BCJ, NRB, PJC, FCS, CZ, PVB, SPB provided critical revision of the manuscript for important intellectual content. SB, BCJ, and SPB provided statistical analysis. SPB supervised the study.

Acknowledgments

The authors thank the staff and participants of the COPDGene study for their important contributions. This study was supported by NIH grants K23 HL133438 and R21EB027891 (to SPB) and an American Thoracic Foundation 2018 Unrestricted Research Grant (to SB). JMR was supported by an NIH grant. MLM was supported by grants from the NIH and the US Department of Veteran Affairs. The COPDGene study is supported by NIH grants NHLBI U01 HL089897 and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprising AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Novartis, and Sunovion.

Address correspondence to: Surya P. Bhatt, University of Alabama at Birmingham, Division of Pulmonary, Allergy and Critical Care Medicine, THT 422, 1720, 2nd Avenue South, Birmingham, Alabama 35294, USA. Phone: 205.934.5555; Email: sbhatt@uabmc.edu.

1. US Burden of Disease Collaborators, et al. The State of US Health, 1990-2016: Burden of Diseases, Injuries, and Risk Factors Among US States. *JAMA*. 2018;319(14):1444–1472.
2. Hogg JC, et al. The nature of small-airway obstruction in chronic obstructive pulmonary disease. *N Engl J Med*. 2004;350(26):2645–2653.
3. McDonough JE, et al. Small-airway obstruction and emphysema in chronic obstructive pulmonary disease. *N Engl J Med*. 2011;365(17):1567–1575.
4. Regan EA, et al. Clinical and radiologic disease in smokers with normal spirometry. *JAMA Intern Med*. 2015;175(9):1539–1549.
5. Woodruff PG, et al. Clinical significance of symptoms in smokers with preserved pulmonary function. *N Engl J Med*. 2016;374(19):1811–1821.
6. Elbehairy AF, Parraga G, Webb KA, Neder JA, O'Donnell DE, Canadian Respiratory Research Network (CRRN). Mild chronic obstructive pulmonary disease: why spirometry is not sufficient! *Expert Rev Respir Med*. 2017;11(7):549–563.
7. Bhatt SP, et al. Imaging Advances in Chronic Obstructive Pulmonary Disease. Insights from the Genetic Epidemiology of Chronic Obstructive Pulmonary Disease (COPDGene) Study. *Am J Respir Crit Care Med*. 2019;199(3):286–301.
8. Coxson HO. Quantitative computed tomography assessment of airway wall dimensions: current status and potential applications for phenotyping chronic obstructive pulmonary disease. *Proc Am Thorac Soc*. 2008;5(9):940–945.
9. Coxson HO, Rogers RM. Quantitative computed tomography of chronic obstructive pulmonary disease. *Acad Radiol*. 2005;12(11):1457–1463.
10. Gietema HA, Edwards LD, Coxson HO, Bakke PS, ECLIPSE Investigators. Impact of emphysema and airway wall thickness on quality of life in smoking-related COPD. *Respir Med*. 2013;107(8):1201–1209.
11. Grydeland TB, et al. Quantitative computed tomography measures of emphysema and airway wall thickness are related to respiratory symptoms. *Am J Respir Crit Care Med*. 2010;181(4):353–359.
12. Han MK, et al. Association between emphysema and chronic obstructive pulmonary disease outcomes in the COPDGene and SPIROMICS Cohorts: A Post Hoc Analysis of Two Clinical Trials. *Am J Respir Crit Care Med*. 2018;198(2):265–267.
13. Johannessen A, et al. Mortality by level of emphysema and airway wall thickness. *Am J Respir Crit Care Med*. 2013;187(6):602–608.
14. Li H, Liu C, Zhang Y, Xiao W. The concave shape of the forced expiratory flow-volume curve in 3 seconds is a practical surrogate of FEV_1/FVC for the diagnosis of airway limitation in inadequate spirometry. *Respir Care*. 2017;62(3):363–369.

15. Nève V, Edmé JL, Baquet G, Matran R. Reference ranges for shape indices of the flow-volume loop of healthy children. *Pediatr Pulmonol*. 2015;50(10):1017–1024.
16. O'Donnell CR, Rose RM. The flow-ratio index. An approach for measuring the influence of age and cigarette smoking on maximum expiratory flow-volume curve configuration. *Chest*. 1990;98(3):643–646.
17. O'Donnell CR, Sneddon SL, Schenker M, Garshick E, Speizer FE, Mead J. Accuracy of spirometric and flow-volume indices obtained by digitizing volume-time tracings. *Am Rev Respir Dis*. 1987;136(1):108–112.
18. Ohwada A, Takahashi K. Concave pattern of a maximal expiratory flow-volume curve: a sign of airflow limitation in adult bronchial asthma. *Pulm Med*. 2012;2012:797495.
19. Zheng CJ, Adams AB, McGrail MP, Marini JJ, Greaves IA. A proposed curvilinearity index for quantifying airflow obstruction. *Respir Care*. 2006;51(1):40–45.
20. Wang W, Xie M, Dou S, Cui L, Xiao W. Computer quantification of "angle of collapse" on maximum expiratory flow volume curve for diagnosing asthma-COPD overlap syndrome. *Int J Chron Obstruct Pulmon Dis*. 2016;11:3015–3022.
21. Bergin C, et al. The diagnosis of emphysema. A computed tomographic-pathologic correlation. *Am Rev Respir Dis*. 1986;133(4):541–546.
22. Bodduluri S, et al. Airway fractal dimension predicts respiratory morbidity and mortality in COPD. *J Clin Invest*. 2018;128(12):5374–5382.
23. Koo HK, et al. Small airways disease in mild and moderate chronic obstructive pulmonary disease: a cross-sectional study. *Lancet Respir Med*. 2018;6(8):591–602.
24. Kapp MC, Schachter EN, Beck GJ, Maunder LR, Witek TJ. The shape of the maximum expiratory flow volume curve. *Chest*. 1988;94(4):799–806.
25. Omland O, Sigsgaard T, Pedersen OF, Miller MR. The shape of the maximum expiratory flow-volume curve reflects exposure in farming. *Ann Agric Environ Med*. 2000;7(2):71–78.
26. Orlowska K, Grebska E. [Examination of the accuracy of spirometric parameters: comparison of FMF (forced midexpiratory flow) and FEV 1 (forced expiratory volume)]. *Gruzlica*. 1972;40(5):425–430.
27. Schachter EN, Kapp MC, Maunder LR, Beck G, Witek TJ. Smoking and cotton dust effects in cotton textile workers: an analysis of the shape of the maximum expiratory flow volume curve. *Environ Health Perspect*. 1986;66:145–148.
28. Tien YK, Elliott EA, Mead J. Variability of the configuration of maximum expiratory flow-volume curves. *J Appl Physiol Respir Environ Exerc Physiol*. 1979;46(3):565–570.
29. Varga J, et al. Relation of concavity in the expiratory flow-volume loop to dynamic hyperinflation during exercise in COPD. *Respir Physiol Neurobiol*. 2016;234:79–84.
30. Jones PW, et al. Responder analyses for treatment effects in COPD using the St George's Respiratory Questionnaire. *Chronic Obstr Pulm Dis*. 2017;4(2):124–131.
31. Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: A strong baseline. 2017 International Joint Conference on Neural Networks (IJCNN). doi: 10.1109/IJCNN.2017.7966039. https://arxiv.org/abs/1611.06455. Accessed on June 23, 2020.
32. Dau HA, et al. UCR Time Series Classification Archive. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/. Accessed June 23, 2020.
33. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med*. 2020;172(1):59–60.
34. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. *JAMA Intern Med*. 2019;179(3):293–294.
35. Regan EA, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2010;7(1):32–43.
36. Bhatt SP, et al. Discriminative accuracy of FEV1:FVC thresholds for COPD-related hospitalization and mortality. *JAMA*. 2019;321(24):2438–2447.
37. Vogelmeier CF, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *Am J Respir Crit Care Med*. 2017;195(5):557–582.
38. Wan ES, et al. Epidemiology, genetics, and subtyping of preserved ratio impaired spirometry (PRISm) in COPDGene. *Respir Res*. 2014;15:89.
39. Bhatt SP, et al. New spirometry indices for detecting mild airflow obstruction. *Sci Rep*. 2018;8(1):17484.
40. Bhatt SP, et al. The peak index: spirometry metric for airflow obstruction severity and heterogeneity. *Ann Am Thorac Soc*. 2019;16(8):982–989.
41. Galbán CJ, et al. Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. *Nat Med*. 2012;18(11):1711–1715.
42. Bhatt SP, et al. Association between functional small airway disease and FEV1 decline in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2016;194(2):178–184.
43. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.
44. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.